Published in Journals: AI, Big Data and Cognitive Computing, Fire, GeoHazards and Remote Sensing

Topic Reprint

AI for Natural Disasters Detection, Prediction and Modeling

Volume II

Edited by
Moulay A. Akhloufi and Mozhdeh Shahbazi

mdpi.com/topics

MDPI

AI for Natural Disasters Detection, Prediction and Modeling—Volume II

AI for Natural Disasters Detection, Prediction and Modeling—Volume II

Topic Editors

Moulay A. Akhloufi
Mozhdeh Shahbazi

MDPI

Basel • Beijing • Wuhan • Barcelona • Belgrade • Novi Sad • Cluj • Manchester

Topic Editors

Moulay A. Akhloufi
Dept of Computer Science
Université de Moncton
Moncton
Canada

Mozhdeh Shahbazi
Department of Geomatics
Engineering
University of Calgary
Calgary
Canada

Editorial Office
MDPI AG
Grosspeteranlage 5
4052 Basel, Switzerland

This is a reprint of the Topic, published open access by the journals *AI* (2673-2688), *Big Data and Cognitive Computing* (ISSN 2504-2289), *Fire* (ISSN 2571-6255), *GeoHazards* (ISSN 2624-795X) and *Remote Sensing* (ISSN 2072-4292), freely accessible at: https://www.mdpi.com/topics/XJL0V1XRH2.

For citation purposes, cite each article independently as indicated on the article page online and as indicated below:

Lastname, A.A.; Lastname, B.B. Article Title. *Journal Name* **Year**, *Volume Number*, Page Range.

Volume II
ISBN 978-3-7258-9079-8 (Hbk)
ISBN 978-3-7258-9080-4 (PDF)
https://doi.org/10.3390/books978-3-7258-9080-4

Set
ISBN 978-3-7258-9075-0 (Hbk)
ISBN 978-3-7258-9076-7 (PDF)

Cover image courtesy of Moulay A. Akhloufi via USGS

© 2026 by the authors. Articles in this reprint are Open Access and distributed under the Creative Commons Attribution (CC BY) license. The reprint as a whole is distributed by MDPI under the terms and conditions of the Creative Commons Attribution-NonCommercial-NoDerivs (CC BY-NC-ND) license (https://creativecommons.org/licenses/by-nc-nd/4.0/).

Contents

About the Editors . . . . . **vii**

Preface . . . . . **ix**

Ziying Gu, Shumin Wang, Kangsan Yu, Yuanhao Wang and Xuehua Zhang
UEBNet: A Novel and Compact Instance Segmentation Network for Post-Earthquake Building Assessment Using UAV Imagery
Reprinted from: *Remote Sens.* **2025**, *17*, 3530, https://doi.org/10.3390/rs17213530 . . . . . **1**

Jingjia Luo, Peng Yang and Fan Meng
Probabilistic Forecasting Model for Tropical Cyclone Intensity Based on Diffusion Model
Reprinted from: *Remote Sens.* **2025**, *17*, 3600, https://doi.org/10.3390/rs17213600 . . . . . **24**

Maryam Nasourinia and Kalpdrum Passi
Wildfire Prediction in British Columbia Using Machine Learning and Deep Learning Models: A Data-Driven Framework
Reprinted from: *Big Data Cogn. Comput.* **2025**, *9*, 290, https://doi.org/10.3390/bdcc9110290 . . . **43**

Cao Yang, Zhou Jun, Wen Hongyuan and Wang Gang
TCSN- YOLO: A Small-Target Object Detection Method for Fire Smoke
Reprinted from: *Fire* **2025**, *8*, 466, https://doi.org/10.3390/fire8120466 . . . . . **68**

Yaning Liu, Jing Ren, Jiakun Wang, Shaoda Li, Rui Chen, Dongsheng Zhong, et al.
LENet: A Semantic Segmentation Network for Complex Landforms in Remote Sensing Imagery via Axial Semantic Modeling and Deformation-Aware Compensation
Reprinted from: *Remote Sens.* **2026**, *18*, 59, https://doi.org/10.3390/rs18010059 . . . . . **86**

Jun Wang, Hongdong Fan, Wanbing Tuo and Yiru Ren
A Multi-Channel Convolutional Neural Network Model for Detecting Active Landslides Using Multi-Source Fusion Images
Reprinted from: *Remote Sens.* **2026**, *18*, 126, https://doi.org/10.3390/rs18010126 . . . . . **111**

Dalei Peng, Maoyuan Chen, Yeping Zhou, Pinliang Li, Shihao Xiao, Yuyang Shen, et al.
A Hybrid Physics–Machine Learning Framework for Landslide Susceptibility Assessment with an Improved Non–Landslide Sampling Strategy
Reprinted from: *Remote Sens.* **2026**, *18*, 408, https://doi.org/10.3390/rs18030408 . . . . . **137**

Gihwan Jung, Tae-Hyuk Ahn and Byungseok Min
Early Wildfire Smoke Detection with a Multi-Resolution Framework and Two-Stage Classification Pipeline
Reprinted from: *Fire* **2026**, *9*, 92, https://doi.org/10.3390/fire9020092 . . . . . **161**

Chenyu Ma, Siquan Yang, Jing Cui, Qiang Li, Qichao Yao, De Zhang, et al.
Applying an Interpretable Deep Learning Model to Identify Wildfire-Prone Areas in Southwest China
Reprinted from: *Fire* **2026**, *9*, 107, https://doi.org/10.3390/fire9030107 . . . . . **190**

Zengzhou Hao, Xin Li, Qiankun Zhu, Yunzhou Li, Zhihua Mao, Jianyu Chen and Delu Pan
AMFA-DeepLab: An Improved Lightweight DeepLabV3+ Adaptive Multi-Statistic Fusion Attention Network for Sea Ice Segmentation in GaoFen-1 Images
Reprinted from: *Remote Sens.* **2026**, *18*, 783, https://doi.org/10.3390/rs18050783 . . . . . **211**

Adnan Ahmed Abi Sen, Fares Hamad Aljohani, Nour Mahmoud Bahbouh, Adel Ben Mnaouer, Omar Tayan and Ahmad. B. Alkhodre
A Context-Aware Flood Warning Framework Integrating Ensemble Learning and LLMs
Reprinted from: *GeoHazards* **2026**, *7*, 35, https://doi.org/10.3390/geohazards7010035 . . . . . **232**

Junyi Jiang, Dong Li, Qingyi Yang, Zhenhua Zhang, Lei Wang, Wenru Zhao and Mingliang Chen
An Intelligent Evaluation Method for Slope Stability Based on a Database Integrating Real Cases and Numerical Simulations
Reprinted from: *Big Data Cogn. Comput.* **2026**, *10*, 87, https://doi.org/10.3390/bdcc10030087 . . **257**

Clara Ochoa, Magí Franquesa, Marcos Rodrigues and Emilio Chuvieco
Inferring Wildfire Ignition Causes in Spain Using Machine Learning and Explainable AI
Reprinted from: *Fire* **2026**, *9*, 138, https://doi.org/10.3390/fire9040138 . . . . . **278**

Tao Guo, Hua Zhang, Tao Song and Shiqiu Peng
Interval-Based Tropical Cyclone Intensity Forecasting with Spatiotemporal Transformers
Reprinted from: *Remote Sens.* **2026**, *18*, 1069, https://doi.org/10.3390/rs18071069 . . . . . **298**

Guoxiu Zhang, Qiangyu Zeng, Fugui Zhang, Hao Wang and Tiantian Yu
TDA-DARKNet: A Deep Learning Model Based on Dual-Polarization Radar Data for Tornado Detection
Reprinted from: *Remote Sens.* **2026**, *18*, 1124, https://doi.org/10.3390/rs18081124 . . . . . **321**

Sara Mostofi, Fatih Yesevi Okur, Ahmet Can Altunişik and Ertugrul Taciroğlu
Deep Unsupervised Learning for Indoor Fire Detection Using Wi-Fi Signals
Reprinted from: *Fire* **2026**, *9*, 189, https://doi.org/10.3390/fire9050189 . . . . . **343**

Shruti Kshirsagar, Bharath Chandra , Unaza Tallal, Rajiv Bagai and Atri Dutta
Geographic Bias Analysis and Cross-Domain Generalization in Deep Learning-Based Building Damage Assessment
Reprinted from: *Remote Sens.* **2026**, *18*, 1529, https://doi.org/10.3390/rs18101529 . . . . . **363**

Boksoo Choi and Gye-Young Kim
Nationwide Daily Wildfire Occurrence Prediction Using Time Proxy Variables and the Canadian Fire Weather Index (FWI)
Reprinted from: *Fire* **2026**, *9*, 217, https://doi.org/10.3390/fire9060217 . . . . . **385**

Yongle Yan, Yichao Zhao and Jiuwu Hui
Research on Gas Concentration Prediction Method Based on Decoupling of Temporal Feature and Dynamic Relationship Reconstruction
Reprinted from: *Fire* **2026**, *9*, 267, https://doi.org/10.3390/fire9070267 . . . . . **411**

About the Editors

Moulay A. Akhloufi

Moulay A. Akhloufi, PhD, holds a PhD and M.Sc. in electrical engineering. He is a Professor in the Department of Computer Science at the Université de Moncton and Director of the PRIME (Perception, Robotics and Intelligent Machines) Research Group. He also holds the AI for Health Research Chair. His expertise spans artificial intelligence, deep learning, computer vision, medical image analysis, and robot vision, with applications in healthcare, environmental monitoring, smart cities, energy, and industrial systems. Professor Akhloufi has led numerous multidisciplinary research initiatives in collaboration with academic institutions, government agencies, hospitals, and industry partners. He has secured millions of dollars in competitive research funding and has supervised and mentored a large number of graduate students and research trainees. Recognized internationally for his scientific contributions, he has been listed for multiple consecutive years among the top 2% of the world's most influential scientists and ranks among the leading researchers in artificial intelligence and deep learning. He is a Senior Member of the Institute of Electrical and Electronics Engineers (IEEE) and a Member of SPIE, the International Society for Optics and Photonics. Through his research, teaching, and community engagement activities, Professor Akhloufi continues to advance the development and adoption of responsible, innovative, and impactful AI technologies in Canada and around the world.

Mozhdeh Shahbazi

Mozhdeh Shahbazi (PhD, PEng) holds a BSc degree in survey engineering, MSc in photogrammetry engineering, and PhD in remote sensing. Her expertise is in the areas of computer vision, deep learning, photogrammetry, sensor-integration and self-calibration, and robot vision. Mozhdeh Shahbazi, is a Physical Scientist Expert at Natural Resources Canada specializing in computer vision, deep learning, remote sensing, photogrammetry, and robotics. She has held academic and research leadership positions at the University of Calgary, York University, the Centre de géomatique du Québec, and Genetec Inc. Throughout her career, she has secured more than $5 million in R&D funding, established multiple research laboratories, and trained over 30 highly qualified personnel. Her work focuses on developing advanced AI and computer vision solutions for intelligent surveillance, geospatial analytics, and natural resource monitoring. She is also an active leader in the remote sensing community, serving in senior roles with national and international scientific societies, journals, and STEM outreach initiatives.

Preface

This Reprint presents a collection of recent advances and innovative research contributions related to the development and application of artificial intelligence approaches for addressing complex challenges associated with natural hazards and environmental risk management. The aim of this Reprint is to highlight emerging methodologies, data-driven solutions, and intelligent systems that support improved understanding, prediction, monitoring, and decision-making in the context of natural disasters.

The motivation for preparing this Reprint comes from the growing need for reliable and efficient technologies capable of supporting communities, researchers, and decision-makers in facing increasingly complex environmental challenges. Advances in artificial intelligence, machine learning, remote sensing, data analytics, and simulation technologies provide new opportunities to enhance disaster preparedness, response strategies, and resilience planning.

This Reprint is intended for researchers, engineers, practitioners, policy-makers, and professionals working in artificial intelligence, environmental science, disaster management, and related interdisciplinary fields. It aims to provide valuable insights into current developments, encourage knowledge exchange, and inspire future research directions to move toward safer and more sustainable societies.

Moulay A. Akhloufi and Mozhdeh Shahbazi
Topic Editors

remote sensing

MDPI

Article

UEBNet: A Novel and Compact Instance Segmentation Network for Post-Earthquake Building Assessment Using UAV Imagery

Ziying Gu [1], Shumin Wang [1,*], Kangsan Yu [2], Yuanhao Wang [1] and Xuehua Zhang [3]

[1] Institute of Earthquake Forecasting, China Earthquake Administration, Beijing 100036, China; guziying@ief.ac.cn (Z.G.); wangyh@ief.ac.cn (Y.W.)
[2] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China; yukangsan@whu.edu.cn
[3] National Earthquake Response Support Service, Beijing 100049, China; zhangxhua_nerss@163.com
* Correspondence: wangsm@ief.ac.cn

Highlights:

What are the main findings?

- UEBNet, a novel instance segmentation model integrating DSCBAM, MFT, and AMSI modules, improves the accuracy of damaged building extraction from UAV imagery in complex post-earthquake scenarios.
- The model was rigorously validated using UAV data from the 2025 Tingri and 2023 Jishishan earthquakes, demonstrating superior performance and robustness against background clutter, scale variation, and irregular damage contours.

What is the implication of the main finding?

- Provides a reliable, high-precision deep learning solution for rapid post-earthquake building damage assessment, supporting efficient emergency response and disaster relief decision-making.
- Offers a scalable architectural framework and module design philosophy applicable to other complex remote sensing instance segmentation tasks beyond seismic damage evaluation.

Abstract

Unmanned aerial vehicle (UAV) remote sensing is critical in assessing post-earthquake building damage. However, intelligent disaster assessment via remote sensing faces formidable challenges from complex backgrounds, substantial scale variations in targets, and diverse spatial disaster dynamics. To address these issues, we propose UEBNet, a high-precision post-earthquake building instance segmentation model that systematically enhances damage recognition by integrating three key modules. Firstly, the Depthwise Separable Convolutional Block Attention Module suppresses background noise that visually resembles damaged structures. This is achieved by expanding the receptive field using multi-scale pooling and dilated convolutions. Secondly, the Multi-feature Fusion Module generates scale-robust feature representations for damaged buildings with significant size differences by processing feature streams from different receptive fields in parallel. Finally, the Adaptive Multi-Scale Interaction Module accurately reconstructs the irregular contours of damaged buildings through an advanced feature alignment mechanism. Extensive experiments were conducted using UAV imagery collected after the Ms 6.8 earthquake in Tingri County, Tibet Autonomous Region, China, on 7 January 2025, and the Ms 6.2 earthquake in Jishishan County, Gansu Province, China, on 18 December 2023. Results indicate that UEBNet enhances segmentation mean Average Precision (mAP^{seg}) and bounding box mean Average Precision (mAP^{box}) by 3.09% and 2.20%, respectively, with

equivalent improvements of 2.65% in F1-score and 1.54% in overall accuracy, outperforming state-of-the-art instance segmentation models. These results demonstrate the effectiveness and reliability of UEBNet in accurately segmenting earthquake-damaged buildings in complex post-disaster scenarios, offering valuable support for emergency response and disaster relief.

Keywords: remote sensing; earthquake disaster; unmanned aerial vehicle (UAV) imagery; instance segmentation

1. Introduction

Earthquake-induced building damage is the primary cause of economic losses and casualties [1]. The swift and accurate assessment of building damage post-earthquake is essential for emergency response and post-disaster reconstruction [2,3]. Due to the harsh post-earthquake disaster site environment, limited communication capabilities, and complex transportation conditions, traditional field survey methods used to map building damage areas are dangerous, inefficient, and costly. The advancement of high-resolution remote sensing technology has led to a growing utilisation of remote sensing imagery in seismic disaster assessment [4,5]. UAV remote sensing technology provides benefits such as high resolution, cheap cost, and enhanced mobility, attracting growing interest from researchers in disaster information collection and earthquake damage assessment [6–12].

Methods for extracting building information from post-earthquake drone imagery generally include visual interpretation, machine learning, and deep learning. The swift advancement of deep learning in computing has rendered deep learning-based remote sensing image information extraction a focal point of research and a prospective trend among experts [13–15]. For example, networks such as Mask R-CNN [16], U-Net [17], and Mask2Former [18] have been used by researchers to extract building damage data from post-earthquake UAV images autonomously. Duarte et al. [19] integrated three photos of varying resolutions into a CNN for the classification of building damage. This model is limited to classifying damaged buildings from images and cannot locate individual damaged targets. Xiong et al. [20] proposed a building picture segmentation approach utilising a 3D building model as a georeference, employing a CNN model based on VGGNet to enhance the accuracy of earthquake damage assessment for structures. In a case study of the old Beichuan town following the 2008 Wenchuan Ms 8.0 earthquake, the method achieved an 89.39% damage identification accuracy. Khankehizadeh et al. [21] proposed a new WETUM method for producing reliable building damage maps, attaining a Damage Detection Rate of 78.26% through the integration of spectral and geometric information. This method holds significant potential and application value for building damage detection in newly emerging earthquake-affected areas. Nonetheless, current object detection techniques are limited to providing bounding boxes for damaged structures [22] and are incapable of delineating their intricate features. To obtain fine-grained contours, Gomroki et al. [23] proposed an encoder–decoder semantic segmentation network based on post-earthquake remote sensing images to extract target contours. This method was experimentally validated on three satellite image datasets from the 2023 Turkey earthquake, the 2023 Morocco earthquake, and the 2023 Libya flood, all of which yielded good results. Xie et al. [10] designed a local–global context attention module and a feature fusion module utilising self-attention based on the U-Net network. When validated on remote sensing data of building damage from the 2010 Haiti earthquake, the IoU increased by 7.39%.

Due to the strong randomness and spatial heterogeneity of building damage in drone imagery [24], traditional semantic segmentation methods often struggle to maintain consistency in building instances, potentially dividing the same building into multiple damage categories, leading to inaccurate identification of individual building boundaries. As a result, recent research on post-earthquake building extraction has increasingly shifted towards instance segmentation methods [25,26] to improve boundary precision and category accuracy. Zhan et al. [27] improved the classic Mask R-CNN [16] to extract post-disaster damaged buildings. This method classifies damaged buildings into four categories: undamaged, slightly damaged, severely damaged, and collapsed. It achieves an accuracy rate of 92% in identifying severely damaged buildings, with an overall classification accuracy rate of roughly 88% across the four damage categories. However, as image resolution increases, training complexity grows exponentially. To mitigate computational overhead in building instance segmentation for high-resolution remote sensing images, Shen et al. [28] proposed DCT-Mask based on Mask R-CNN. This network uses a 300-dimensional Discrete Cosine Transform (DCT) vector to represent a 128×128 mask, achieving significant performance improvements. Nevertheless, each pixel value in the mask is determined by the entire DCT vector, and modifying part of the vector affects the global structure, potentially disrupting correctly segmented regions. To weaken this global coupling and further optimise segmentation boundaries, Wen et al. [29] proposed PatchDCT. This method disaggregates the mask derived from the DCT vector into several patches and enhances each patch utilising a specified classifier and regressor. The classifier identifies mixed, foreground, and background patches, and can rectify misclassified foreground and background patches from the initial DCT approach. Conversely, the regressor facilitates DCT vector prediction for mixed patches, hence enhancing the segmentation quality of boundary points. However, the aforementioned methods lack a targeted multi-scale feature fusion mechanism, leading to coarse boundaries during decoding and a significant performance decline in the extremely complex post-earthquake scenario.

For complex, unstructured post-earthquake scenarios characterised by irregular building boundaries and low visual distinction between rubble, debris, and damaged structures, we propose UEBNet, a novel instance segmentation framework with three key contributions:

(1) At the feature encoding stage, a Depthwise Separable Convolutional Block Attention Module (DSCBAM) is designed to enhance the ability of the top-level features of the feature pyramid network (FPN) to represent key regions. This module combines multi-scale pooling with depthwise separable dilated convolutions to improve multi-scale semantic modelling capabilities, expand the receptive field, and fundamentally suppress strong background noise interference from bare fields, garbage piles, etc., thereby addressing the challenge of 'accurate perception' in post-earthquake scenes.

(2) In the feature fusion layer, the Multi-feature Fusion Module (MFT) is constructed to fuse semantic features from different levels, enhancing the model's perception of buildings with significant scale differences and incomplete shapes. This module ensures feature consistency and expressiveness, particularly excelling in boundary localization of small objects or damaged areas.

(3) At the decoding and output end, the Adaptive Multi-Scale Interaction Module (AMSI) is innovatively developed. This module addresses the common issue of coarse mask semantic bias in existing methods through unique spatial decomposition, cross-attention, and selective fusion mechanisms. It precisely aligns and reinforces decoded features, enabling the perfect capture of the complete contours of irregular buildings.

2. Study Area and Dataset

2.1. Study Area

According to the China Earthquake Networks Center, an earthquake occurred at 9:05 a.m. Beijing Time on 7 January 2025, at Tingri County, Shigatse City, Tibet Autonomous Region, China (28.50°N, 87.45°E). The earthquake had a magnitude of Ms 6.8 and a focal depth of 10 km, making it the largest earthquake in the region in the past five years. The maximum intensity of this earthquake was IX, covering a region of 411 km^2. The area with an intensity above VI covered approximately 23,986 km^2 [30]. The earthquake resulted in 126 fatalities and affected 61,500 people.

In the rural areas near the epicentre, the majority of buildings were self-built houses. Constrained by local conditions and traditional living customs, these houses were typically constructed using locally sourced materials, predominantly wood-framed with mud brick or stone infill. They lacked systematic seismic design, often featuring irrational structural systems, which led to poor seismic performance [31]. In the seismic intensity zones IX and VIII, the proportions of earth-wood houses with poor seismic resistance were 97% and 88%, respectively. The overall collapse rate in intensity zone IX exceeded 60%. Consequently, the damage to housing was extremely severe. According to statistics, a total of 27,248 houses were damaged, of which 3612 collapsed [32]. Figure 1 depicts the location of the study region used in this paper.

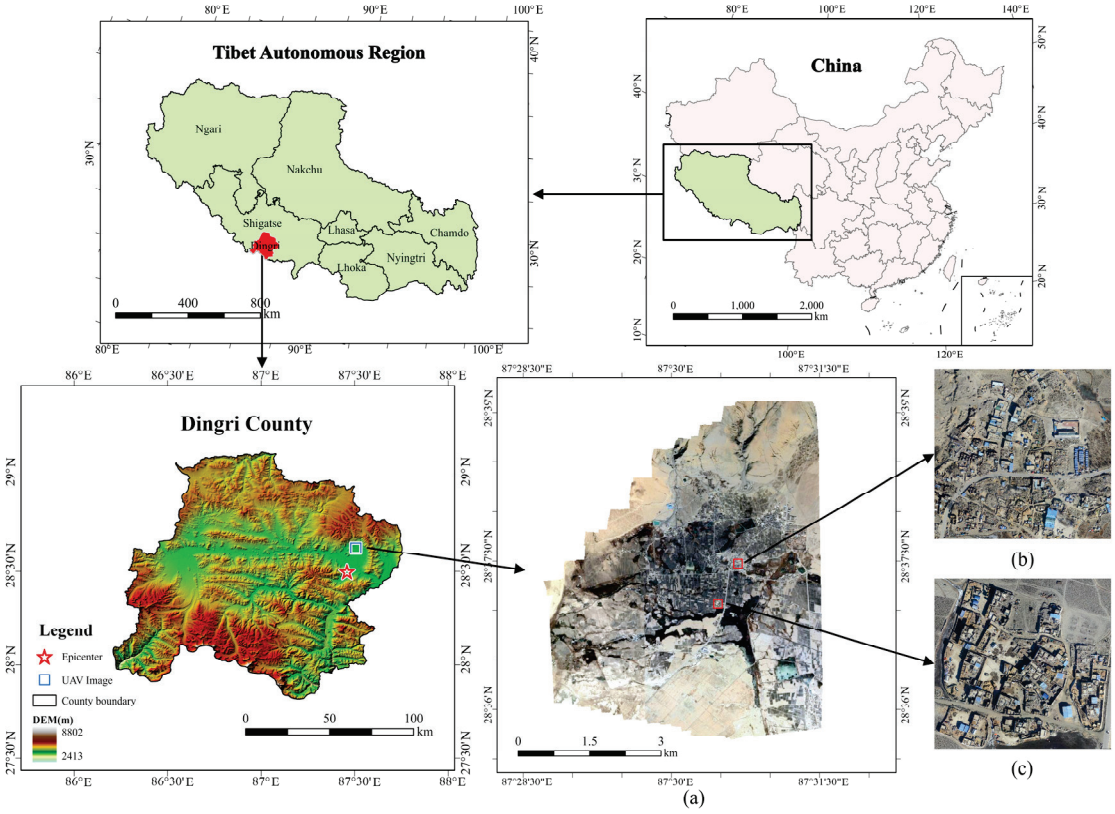


Figure 1. The study area and post-earthquake UAV orthophotos of the Tingri earthquake. (**a**) UAV image. (**b**) Chajiang village. (**c**) Tonglai village.

In the study area, we used a Feima UAV (Feima Robotics, Shenzhen, China) equipped with a SONY ILCE-7RM4A camera (Sony Group Corporation, Tokyo, Japan), which has a focal length of 20 mm. The images of the area near Changsuo Township in Tingri County were acquired on 20 January 2025, from a flight altitude of 500 m. Following processing, a UAV orthoimage with a resolution of 5 cm was produced (as shown in Figure 1a). The

red box in Figure 1a indicates the region cropped for model inference to test the model's applicability, which includes Chajiang Village (Figure 1b) and Tonglai Village (Figure 1c). The remaining areas were used for model training, validation, and evaluation.

2.2. Post-Earthquake Buildings Dataset

To construct the instance segmentation dataset, this study used the professional software ArcGIS Pro (v3.0) to manually annotate the location and outlines of post-earthquake buildings. To ensure dataset quality, the UAV images were segmented into 640 × 640 pixel, four-band patches. As many of these images contained only background and no foreground information, they could not be considered valid samples and were therefore discarded. Finally, the annotated samples were converted into the COCO dataset format.

In total, we annotated 790 images, comprising 1929 damaged buildings and 199 intact buildings. Due to the significant difference in sample sizes between classes, focal loss was adopted in the model to balance the categories. Considering the small scale of the dataset, we employed K-fold cross-validation [33,34]. The training set consisted of 553 images, the validation set included 158 images, and the test set contained 79 images. Detailed information about the post-earthquake building dataset is provided in Table 1. Figure 2 shows sample instances from the dataset.

Table 1. Specifications of the sub-datasets employed for training, validating, and testing.

Split	Images	Samples		
		Damaged	Intact	Sum
Training	553	1386	153	1539
Validation	158	356	26	382
Testing	79	187	20	207
Sum	790	1929	199	2128

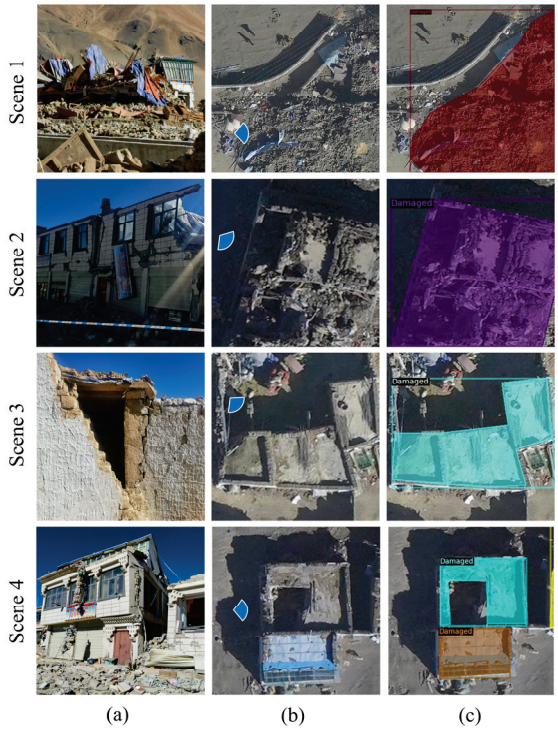


Figure 2. Example of Tingri earthquake data samples: (**a**) field survey photos; (**b**) UAV images, where the observation location's viewing angle is indicated by blue fan shapes; (**c**) sample labels, with different mask colours representing different instances.

3. Methodology

3.1. Overview of UEBNet

As shown in Figure 3, the overall framework of UEBNet consists of three core components: a backbone network, an enhanced feature pyramid neck, and a prediction regression head that incorporates multi-scale feature fusion.

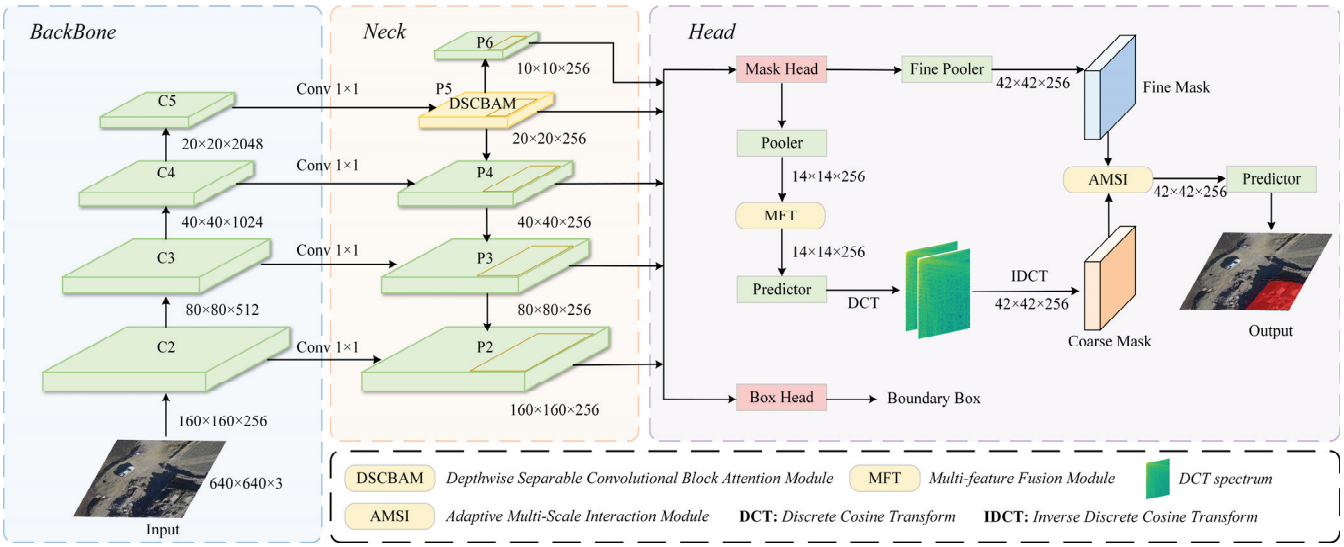


Figure 3. The overall framework of UEBNet. C2–C5 and P2–P6 represent multi-level feature maps.

The input image is first fed into a standard ResNet backbone network for basic feature extraction, and then enters the FPN that has been specifically enhanced to handle post-earthquake scenes. At the top layer of the FPN, we embed the DSCBAM to suppress background noise and enhance key semantic information at an early stage.

Then, the multi-level feature maps generated by the FPN are used for the Region Proposal Network (RPN). For each Region of Interest (RoI), the model not only extracts its corresponding features but also processes them using the MFT. The MFT generates a fusion feature that is more robust to changes in object size by parallel processing multiple scale versions of the feature.

Next, the fused features are fed into the mask regression branch. This branch employs a coarse-to-fine prediction strategy. First, the features output by MFT are subjected to DCT transformation, and the inverse transformation of the first 300 DCT vectors is taken to obtain a coarse-grained global mask. To address the semantic bias issues arising from the direct fusion of coarse- and fine-grained features, we introduce the AMSI. Unlike traditional methods that rely on simple element-wise addition, the AMSI module cross-attention and selectively fuses coarse-grained mask information with high-resolution features from the FPN lower layers, thereby refining the mask.

Finally, the classifier and regressor of PatchDCT are used to predict the mask. Through this series of carefully designed modular components, UEBNet ensures that effective information is maximally retained and utilised at every stage of feature extraction, fusion, and final prediction, thereby achieving high-precision segmentation of irregular, multi-scale architectural targets.

3.2. Depthwise Separable Convolutional Block Attention Module

One of the challenges in disaster remote sensing assessment is how to separate effective target information from visually confusing backgrounds. Especially in post-earthquake scenarios, the features of damaged buildings are often overwhelmed by high-frequency

noise such as debris, dust, and exposed ground surfaces. To enhance the model's capacity to discern features in such complex interference, we present the Convolutional Block Attention Module (CBAM) [35], whose structure is shown in Figure 4a. CBAM learns 'what to focus on' and 'where to look' through serialised channel and spatial attention mechanisms, suppressing interference from irrelevant information. However, while improving accuracy, disaster response also imposes strict requirements on processing speed. The computational overhead of standard CBAM makes it difficult to achieve both. Therefore, we designed a lightweight DSCBAM at the top-level features of FPN. This module replaces ordinary convolutions with computationally cheaper depthwise separable convolutions, reducing the model's parameter count and computational load while preserving feature extraction capabilities to meet the demand for rapid and precise post-disaster assessment.

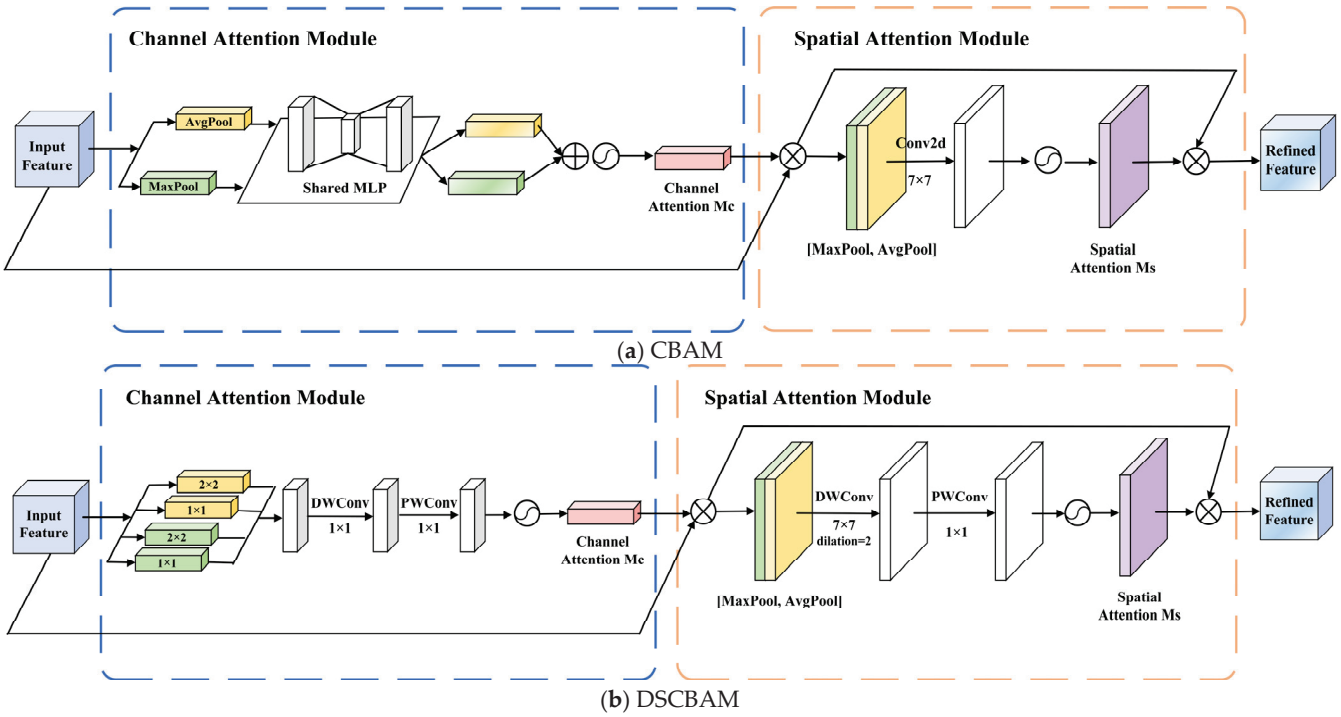


Figure 4. Comparison of network structures between CBAM and DSCBAM.

The DSCBAM module structure is shown in Figure 4b, which includes a Channel Attention and a Spatial Attention. The channel attention module aims to model the interdependencies between channels in feature maps, automatically assigning weights to enhance focus on key features relevant to the current task. To capture richer contextual information, this module employs adaptive pooling operations, including 1×1 and 2×2 adaptive average pooling and max pooling, and fuses the resulting features. To further improve efficiency, the module replaces the traditional multi-layer perceptron (MLP) with a depthwise separable one-dimensional convolution structure, significantly reducing computational complexity. As shown in Figure 5, the depthwise separable convolution [36] first performs depthwise convolution on the input features to extract local features within each channel, then uses pointwise convolution (1×1) to facilitate information exchange between channels. This structure not only simulates the nonlinear transformation and feature fusion functions of the MLP but also markedly decreases the parameter count and computational complexity. After pooling, the feature vectors are reduced in dimension and fused through a linear adapter to generate a 256-dimensional feature vector. The ultimate channel attention weights are derived utilising the sigmoid activation function, and thereafter employed to reconfigure the original feature map.

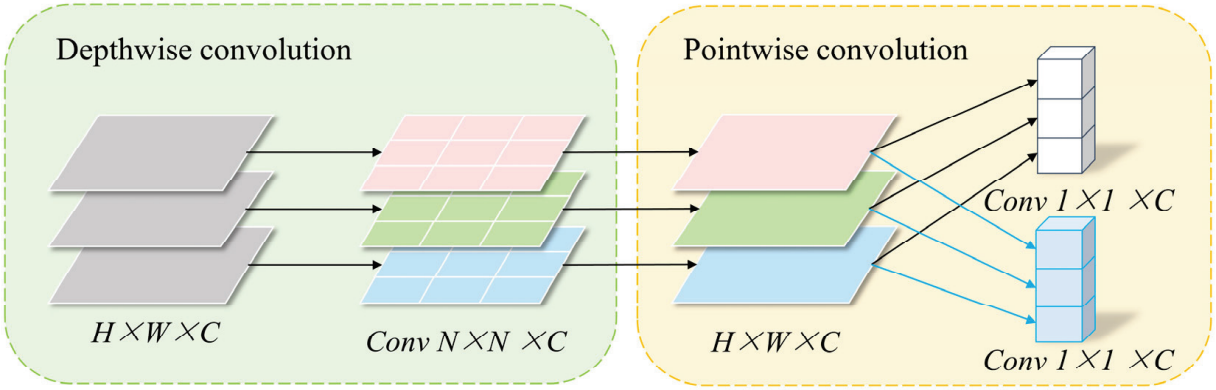


Figure 5. Depthwise separable convolution.

Following channel attention optimisation, the feature maps proceed to the spatial attention module, which is designed to produce a spatial attention map that amplifies the feature response in critical areas. The spatial attention module initially executes global average pooling and max pooling across the channel dimension, subsequently concatenating the two resultant two-dimensional feature maps to create an effective spatial feature descriptor. Unlike traditional convolutions, we employ dilated depthwise separable convolutions, which consist of a 7 × 7 kernel, a depthwise convolution with a dilation rate of (2, 2) [37], and a 1 × 1 pointwise convolution. Dilated convolutions efficiently enlarge the receptive field without augmenting computing demands, allowing the model to acquire more extensive spatial contextual information. The spatial attention map produced by the sigmoid activation function is multiplied element-wise with the input feature map, optimising the feature distribution in the spatial dimension and further augmenting the model's emphasis on critical locations.

3.3. Multi-Feature Fusion Module

The damage patterns of buildings after an earthquake exhibit characteristics of coexisting multiple scales, with the extent of damage varying randomly across buildings of different sizes. The mask head, which uses a single fixed-resolution RoI feature, cannot effectively perceive targets of different scales simultaneously. A receptive field optimised for large-scale debris scenes may overlook details of small-scale earthquake-damaged buildings, and vice versa. This inherent scale limitation results in incomplete semantic representation and reduced localisation accuracy along object boundaries.

To address this issue in post-disaster scenarios, we introduce a lightweight and effective MFT module aimed at enhancing feature quality prior to mask decoding. The MFT module is positioned after the RoIAlign operation as a refinement stage, integrating multi-scale contextual information. Unlike modules that rely on single-scale features, the MFT module simultaneously processes three feature maps obtained by upsampling and downsampling the original RoI-aligned features.

The structure of the MFT module is shown in Figure 6. Specifically, we first upscale the input features (w × h) to a higher spatial resolution (2w × 2h) via bilinear interpolation, while downsampling them by a factor of 0.5 to obtain a low-resolution version (0.5w × 0.5h). The three features of different scales undergo independent nonlinear transformations. The high spatial resolution branch employs max pooling and average pooling to compress feature maps spatially, thereby aggregating global information. Subsequently, to avoid the computational burden of large convolutional kernels, we utilise 1 × 1 convolutions to efficiently facilitate inter-channel feature interaction and dimensionality reduction. The remaining two branches both employ 3 × 3 convolutions to extract local spatial features.

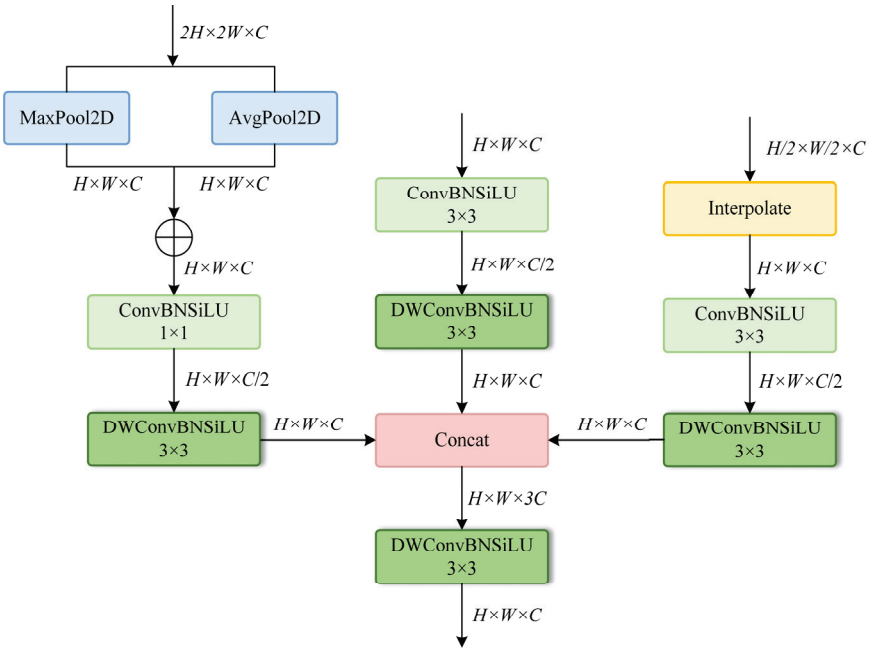


Figure 6. Structure of the MFT module.

To balance feature extraction effectiveness with computational efficiency, each branch internally employs a lightweight depthwise separable convolution. It consists of a 3×3 depthwise convolution (groups = 128) and a 1×1 pointwise convolution, with the depthwise convolution responsible for spatial modelling and the pointwise convolution for channel recovery. After scale normalisation, all three feature maps are concatenated along the channel dimension to form an aggregated feature map with 768 channels (256×3). This aggregated feature map consolidates feature information from different processing paths, offering richer diversity. Finally, the aggregated feature map is fed into an output convolution layer, which consists of a depth-separable convolution, batch normalisation, and a SiLU activation function. Its primary function is to efficiently interact and fuse the multi-source features after concatenation and reduce the channel dimension back to 256. This step not only achieves effective feature integration but also ensures that the feature dimension of the module output matches that of subsequent network layers. This design fuses multi-scale information in a lightweight and modular manner, enabling the network to extract features with higher discriminative power and shape awareness, improving the performance of subsequent instance mask prediction tasks.

3.4. Adaptive Multi-Scale Interaction Module

Accurately delineating the contours of damaged buildings after an earthquake is another challenge in instance segmentation for post-disaster remote sensing assessment. Unlike the clear object boundaries in conventional scenes, the edges of collapsed buildings are structurally chaotic and disordered. In such cases, although UEBNet employs an iterative refinement strategy from coarse to fine to generate masks, the key to determining the final segmentation accuracy lies in how to effectively fuse the initial coarse-grained mask with the fine-grained features from the backbone network. Simple element-wise addition fusion ignores the semantic differences between the two types of features: coarse-grained features emphasise global structure, while fine-grained features emphasise local texture. This may lead to poor feature interaction when handling complex object boundaries or partially collapsed buildings.

To address this issue, we designed an AMSI module to replace simple additive fusion with a more expressive interaction mechanism. Unlike methods that rely on actual

frequency transformations, AMSI simulates multi-scale feature responses by applying convolutional kernels with different receptive fields. Specifically, as shown in Figure 7a, the Spatial Convolution Module (SCM) decomposes fine-grained feature maps into two parallel branches: one branch uses small convolutional kernels to emphasise local textures, while the other branch uses dilated convolutions to capture broader semantic context. Although no explicit frequency transformation is performed, this decomposition simulates the behaviour of high-frequency and low-frequency responses in the spatial domain. Concurrently, lightweight 1×1 convolutions convert the coarse-grained mask features obtained via DCT into query embeddings. Two cross-attention branches are then applied, enabling the coarse-grained features to query the two fine-grained branches, thereby achieving adaptive alignment between structure and detail. Finally, the Selective Feature Fusion (SFF) module dynamically merges the two attention outputs based on learned spatial attention weights. The fused features are added back to the fine-grained features via residual connections and passed to the downstream prediction layer. The SFF module structure is shown in Figure 7b. The AMSI module design achieves content-aware multi-level fusion, respecting the complementarity between structure and detail, enhancing the model's ability to analyse irregular boundaries and maintain semantic consistency, particularly in complex post-disaster scenarios.

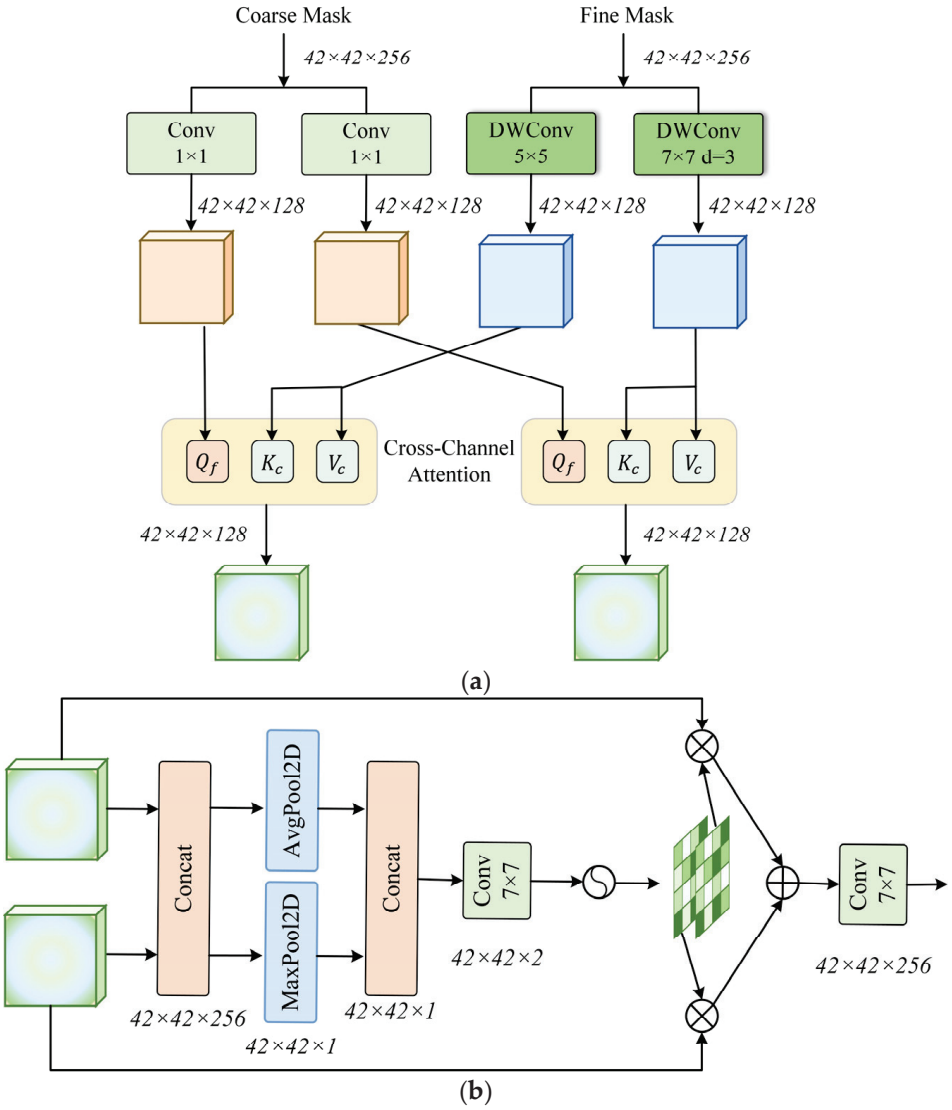


Figure 7. Structure of the AMSI module: (**a**) SCM module; (**b**) SFF module.

3.5. Loss Function

The total loss function used in our framework consists of four components: a focal loss for bounding box classification, a coarse DCT regression loss over the entire mask, a patch-wise classification loss, and a patch-wise DCT regression loss for mixed patches. The specific calculation process is shown in Equations (1)–(4) below.

The overall objective is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{focal} + \lambda_0 \mathcal{L}_{dct,N} + \lambda_1 \left(\mathcal{L}_{cls}^{patch} + \mathcal{L}_{dct}^{patch} \right) \tag{1}$$

where λ_0 and λ_1 are scalar weights, both of which are 0.5 in this article.

The focal loss [38] is used to alleviate class imbalance in building classification and is defined as:

$$\mathcal{L}_{focal} = -\alpha_t (1 - p_t)^{\gamma} \log(p_t) \tag{2}$$

where $p_t \in [0, 1]$ is the predicted probability for the ground-truth class, γ is the focusing parameter (we set $\gamma = 2.0$), α_t and is the class balancing factor (set to 0.25).

The coarse DCT [28] loss measures the difference between predicted and ground-truth DCT vectors of the full-resolution mask:

$$\mathcal{L}_{dct,N} = \frac{1}{N} \sum_{i=1}^{N} R(\hat{V}_i - V_i) \tag{3}$$

where $\hat{V}_i$ and V_i denote the i-th element of predicted and ground-truth DCT vectors, respectively. N is the vector length, and R is a regression function.

The patch classification loss $\mathcal{L}_{cls}^{patch}$ is a cross-entropy loss [39] over three categories. The patch-wise DCT regression loss [29] is applied only to patches classified as mixed:

$$\mathcal{L}_{dct}^{patch} = \frac{1}{N_m} \sum_{i=1}^{N_m} R(\hat{v}_i - v_i) \tag{4}$$

where $\hat{v}_i$ and v_i are the predicted and ground-truth DCT vectors of the i-th mixed patch, and N_m is the total number of mixed patches.

4. Experimental Results and Discussion

4.1. Experimental Configuration

Our model is built using the PyTorch (v1.12.1) framework. The hardware configuration for this experiment is a PC with an Nvidia GeForce RTX 2080 Ti GPU (32 GB), Intel Core i7 CPU, and Ubuntu 18.04. We use the standard partition of training and validation sets in both datasets. Images are cropped to a size of 640 × 640 for training. In the training phase, the random flip is used for data augmentation.

If not specified, we employ ResNet50 with pre-trained weights as the default backbone. The model training process employs an initial learning rate of 0.0005, a batch size of 2, and a total of 120 epochs, optimized using the Adam optimizer.

4.2. Evaluation Metrics

Quantitative evaluation measures are necessary to assess the model's performance. In this study, mean Average Precision (*mAP*) and F1-score are adopted to comprehensively evaluate the detection and segmentation performance of the model. The evaluation metrics can be formulated as follows:

$$P = \frac{TP}{TP + FP} \tag{5}$$

$$R = \frac{TP}{TP + FN} \tag{6}$$

$$F_1 = \frac{2 \times P \times R}{P + R} \tag{7}$$

$$OA = \frac{TP + TN}{TP + FP + TN + FN} \tag{8}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{9}$$

$$AP_{IoU} = \int_0^1 P(R)dR \tag{10}$$

Here, TP and FP represent the true positive and false positive [40], respectively, while TN and FN denote the true negative and false negative. P and R represent Precision and Recall, respectively. The F1-score is the harmonic mean of Precision and Recall. Overall Accuracy (OA) measures the proportion of correctly predicted samples out of the total number of samples. Average Precision (AP) describes the model's prediction performance based on Precision and Recall. At a given Intersection over Union (IoU), when Recall is plotted on the horizontal axis and Precision on the vertical axis, a PR curve is obtained. The value of AP_{IoU} corresponds to the area under the PR curve.

mAP is the average across 10 IoU thresholds, ranging from 0.50 to 0.95 with a step size of 0.05, and can be calculated using Equation (11):

$$mAP = \frac{1}{10} \sum_{IoU=0.5}^{0.95} AP_{IoU} \tag{11}$$

Furthermore, we employ frames per second (FPS) to assess the recognition efficacy of various approaches. The equation is as follows, with T denoting the duration required by the model to infer a single image:

$$FPS = \frac{1}{T} \tag{12}$$

4.3. Quantitative Comparisons with SOTA

To validate the effectiveness and advanced nature of UEBNet for post-earthquake building instance segmentation, we compared it with several state-of-the-art (SOTA) instance segmentation methods. All experiments were conducted using the Tingri post-earthquake building dataset. The methods for comparison included Mask R-CNN [16], DCT-Mask [28], YOLOv8 [41], PointRend [42], YOLACT [43], SOLOv2 [44], PatchDCT [29], Mask2Former [18]. For all models, the backbone network was ResNet50 with pre-trained weights.

The quantitative results, as shown in Table 2, clearly indicate that UEBNet surpasses existing state-of-the-art (SOTA) models across all key evaluation metrics. Specifically, UEBNet realised a segmentation mean Average Precision (mAP^{seg}) of 42.26%, an improvement of 3.09% over PatchDCT. Furthermore, its F1-score and Overall Accuracy (OA) reached 78.74% and 94.15%, respectively, demonstrating effective classification performance. Its superior performance stems from the model's ability to enhance key features and fuse multi-scale features within complex post-disaster scenes.

We also conducted a qualitative comparison on the validation set between the DCT-Mask, PointRend, PatchDCT, and UEBNet models to visually demonstrate their recognition performance on post-earthquake buildings. The visualisation results are shown in Figure 8 and include the following representative scenarios: completely collapsed buildings with

irregular boundaries; a single damaged building; multiple damaged buildings; and a scene with intact buildings.

Table 2. Quantitative results for all models of instance masks. Best results are shown in boldface.

Model	mAP^{box} (%)	$AP^{box}_{0.5}$ (%)	$AP^{box}_{0.75}$ (%)	mAP^{seg} (%)	$AP^{seg}_{0.5}$ (%)	$AP^{seg}_{0.75}$ (%)	F1 (%)	OA (%)	FPS (img/s)
Mask R-CNN	32.23	51.34	32.32	33.56	52.50	37.07	67.14	88.07	30.06
DCT-Mask	35.48	55.46	40.91	35.88	55.83	39.80	72.19	89.91	27.30
YOLOv8-Ins-s	36.27	55.14	39.23	36.86	54.89	39.81	72.48	90.43	**33.91**
PointRend	37.45	55.83	38.92	38.73	56.68	41.58	73.10	90.89	23.36
YOLACT	37.85	58.66	43.78	38.46	56.82	41.89	74.25	90.85	19.35
SOLOv2	38.13	57.19	41.52	38.97	56.46	42.19	73.98	91.51	20.47
PatchDCT	39.74	56.27	42.97	39.17	55.75	42.91	76.09	92.61	16.88
Mask2Former	40.21	57.49	43.56	40.83	57.76	44.28	76.73	92.86	14.52
UEBNet (Ours)	**41.94**	**59.32**	**43.85**	**42.26**	**58.51**	**46.49**	**78.74**	**94.15**	14.04

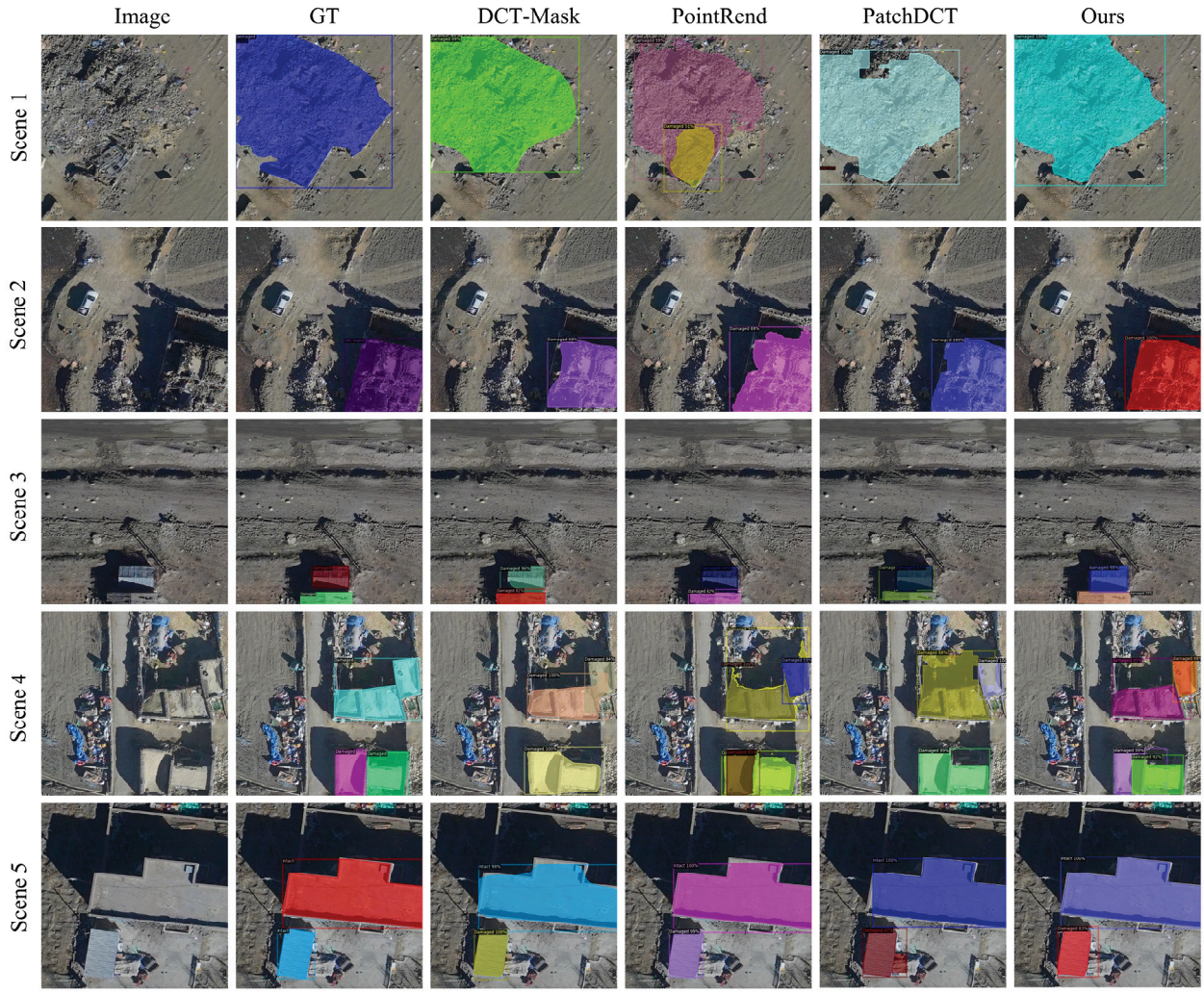


Figure 8. Examples of instance segmentation results from different models on the Tingri earthquake dataset.

All four models were able to accurately locate post-earthquake buildings, but they varied in the precision of their segmentation boundaries and their adaptability to complex scenes. DCT-Mask and PointRend showed poor segmentation capability for multiple buildings, failing to distinguish well between adjacent structures in Scene 4. This exposes their deficiencies in global context awareness and in modelling relationships between instances. Although PointRend uses iterative refinement for points with high uncertainty, its sampling strategy may not effectively decouple the boundaries of different instances in

dense regions. Meanwhile, DCT-Mask, which applies global DCT encoding to the entire instance, struggles to handle the fine gaps between instances. PatchDCT was unable to effectively fuse fine and coarse features, resulting in very poor recognition of damaged building boundaries, especially for the ruins of completely collapsed buildings, with obvious omissions seen in Scene 1.

In contrast, the UEBNet model performed well in localising the boundaries of both damaged and intact buildings. This is due to the synergy between the DSCBAM and AMSI modules. The DSCBAM module uses dilated convolutions to capture a wider area around the target, providing more comprehensive context. Simultaneously, its attention mechanism enables the model to focus on the target region, fundamentally reducing interference from background noise. The AMSI module employs a cross-attention mechanism where coarse structural features actively query fine-grained texture features, aligning structure with detail to precisely reconstruct the irregular contours of collapsed buildings. For example, in Scene 1, where the Ground Truth (GT) indicates that the collapsed building has an extremely complex shape and is similar in color to the bare ground, UEBNet achieved more continuous and precise boundaries. Furthermore, UEBNet also excelled in multi-object recognition and in identifying targets of different sizes. This is because the MFT module creates three parallel streams to process different scales, improving the model's robustness to size variations. In synergy, UEBNet's three innovative modules work to resolve the inherent limitations of traditional methods in complex post-earthquake scenarios.

4.4. Ablation Experimental

To systematically analyse the independent contributions of the three core components within UEBNet, we designed four five sets of ablation studies. By strategically disabling specific functional modules within the complete framework, we were able to quantify the key roles played by DSCBAM, MFT, and AMSI in the post-earthquake building recognition task. Moreover, we have specifically compared the CBAM module with the enhanced DSCBAM module to demonstrate the efficacy and necessity of our improvements. The experimental configurations and results are shown in Table 3, where 'DS' refers to the DSCBAM module, $mAP^{seg}_{dam}(\%)$ is the segmentation precision for the Damaged class, and $mAP^{seg}_{int}(\%)$ is the segmentation precision for the Intact class. The ablation study began with a baseline architecture (Case 1), which represents a simplified version of UEBNet with all core enhancement modules removed. In this configuration, the model retained only the ResNet, a standard FPN, and the basic head of PatchDCT, using the original FPN features directly for candidate region generation and a simple element-wise addition to fuse coarse and fine masks. Subsequently, we progressively 're-activated' UEBNet's built-in components: Case 2 reintegrated the CBAM module at the top layer of the FPN. Case 3 replaced the CBAM module with the DSCBAM module. Building on this, Case 4 further enabled the MFT module within the mask head. Case 5 represented the complete UEBNet framework, with all three core components working in synergy.

Table 3. Results of the ablation experiment.

	Baseline	CBAM	DS	MFT	AMSI	mAP^{box} (%)	mAP^{seg} (%)	mAP^{seg}_{dam} (%)	mAP^{seg}_{int} (%)	F1 (%)	OA (%)	FPS (img/s)
Case 1	√					39.74	39.17	26.17	50.36	76.09	92.61	16.88
Case 2	√	√				40.23 (+0.49)	39.79 (+0.62)	26.75	52.90	76.69	92.64	16.02
Case 3	√		√			40.39 (+0.65)	40.25 (+1.08)	27.56	53.93	77.69	92.65	16.47
Case 4	√		√	√		40.74 (+1.00)	41.63 (+2.46)	28.15	55.04	78.05	93.14	15.88
Case 5	√		√	√	√	41.94 (+2.20)	42.26 (+3.09)	28.23	56.95	78.74	94.15	14.04

In Table 3, the experimental results demonstrate that model performance improved incrementally as modules were added. Moreover, the DSCBAM module achieves higher accuracy and faster FPS than the CBAM module. Figures 9 and 10 visualise the segmentation results of the five experiments, showing the impact of adding different modules to the baseline model. The visual comparison between Case 2 and Case 3 demonstrates that the attention mechanism enables the model to focus more on architecture-related features, thereby enhancing localisation accuracy and reducing false negatives and false positives. The improved DSCBAM module exhibits superior performance. Case 4 shows that the MFT module enhances the model's ability to localise the boundaries of buildings at different scales, proving its role in improving global contextual consistency and local detail precision. The results for Case 5 reflect that the AMSI module respects the complementary nature of structure and detail, enhancing the model's ability to resolve irregular boundaries and maintain semantic consistency. These results prove that each proposed module improves performance to same extent, and that their contributions are not merely additive; their integration further improves performance.

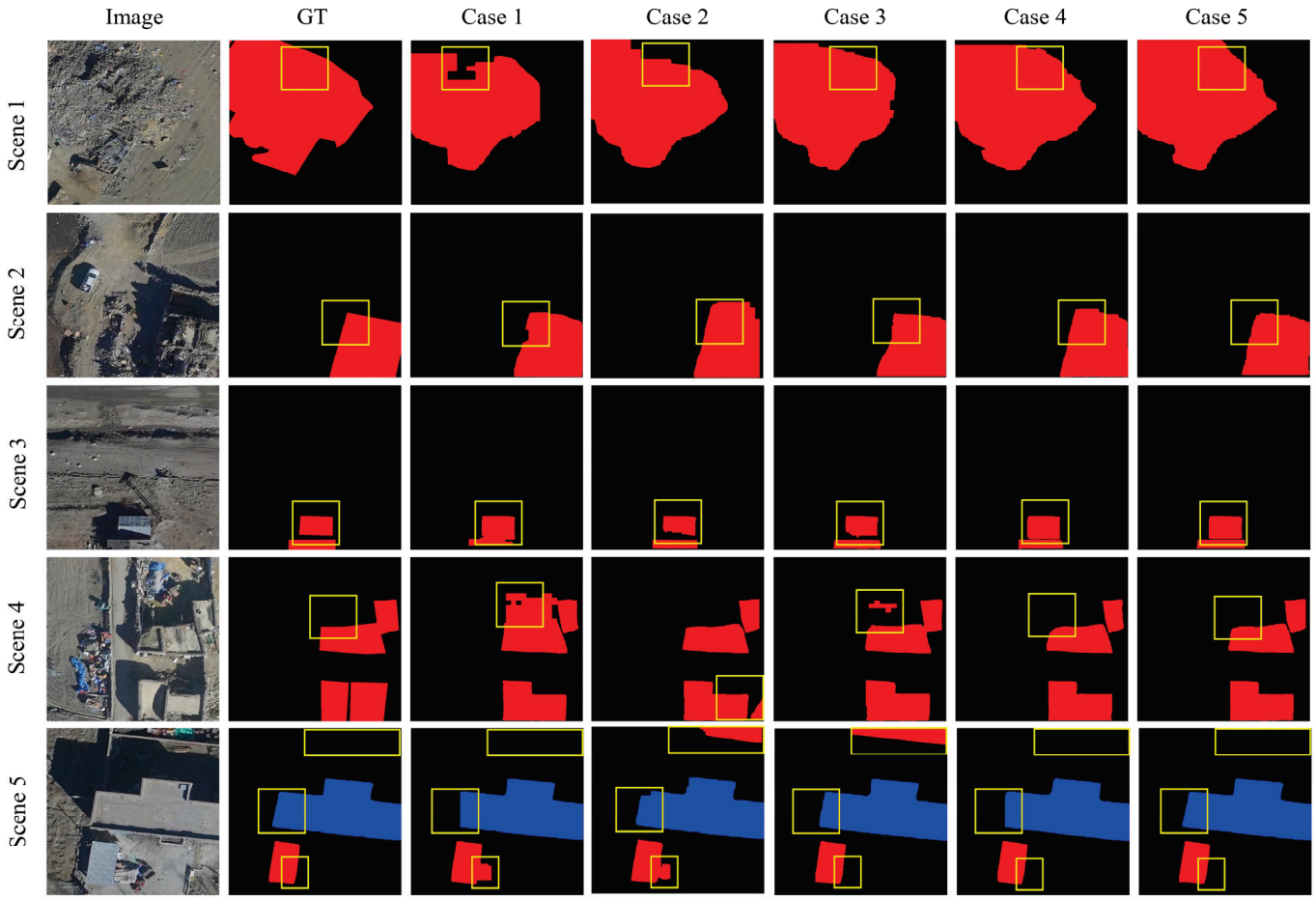


Figure 9. Comparison of segmentation masks in ablation experiments. The red mask represents the Damaged class, the blue mask represents the Intact class, and the yellow boxes indicate areas where segmentation is difficult.

It should be noted that despite a reduction in FPS, the UEBNet model still offers distinct advantages within the specific context of post-disaster assessment. Firstly, disaster response applications demand exceptionally high accuracy and reliability. Even a 2–3% increase in segmentation precision enables the correct evaluation of hundreds of buildings across extensive disaster zones. The costs of false negatives and false positives are extraordinarily high. Therefore, even at the expense of increased computational complexity,

pursuing greater accuracy remains a worthwhile scientific endeavour in this field. Secondly, the reliability of the UEBNet model in complex scenarios is of greater importance. Simplistic models falter in challenging scenarios, such as precisely segmenting highly irregular collapsed boundaries or separating densely clustered structures. The UEBNet, incorporating three innovative modules, partially resolves this dilemma. However, it cannot be denied that model lightweighting facilitates rapid deployment and extreme rescue operations, making it a crucial area of future research.

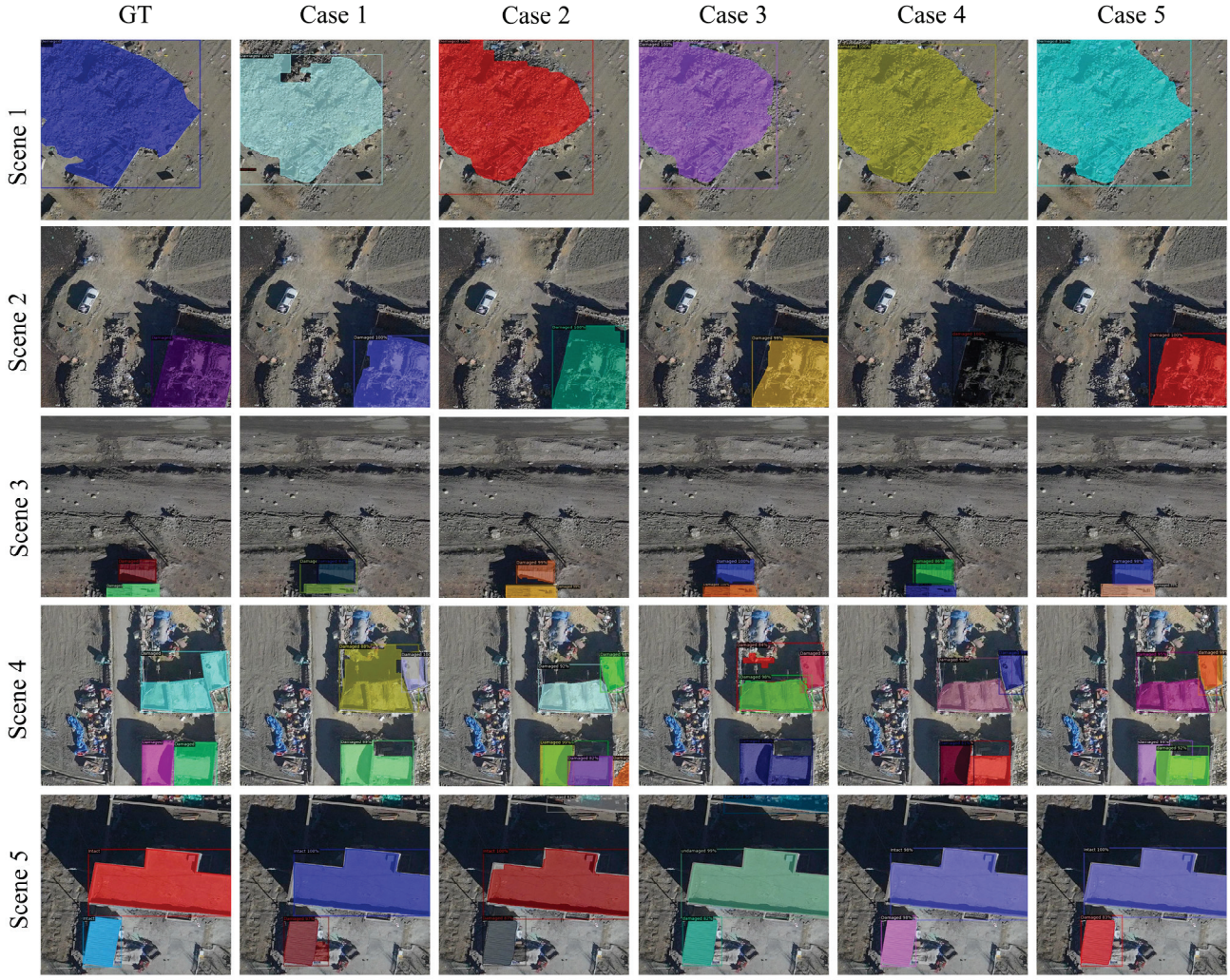


Figure 10. Example of segmentation results from the ablation experiment.

4.5. Applicability of UEBNet Model

To evaluate the practical application performance of the UEBNet model, we conducted inference tests on the regions shown in Figure 1b,c from Section 2.1. As illustrated in Figure 11, UEBNet successfully detected and segmented the various types of buildings within the areas, with the boundaries showing a high degree of conformity to the actual outlines. The classification results were also consistent with those from visual interpretation, exhibiting low rates of misclassification and omission. These results fully confirm that the UEBNet model possesses good accuracy and robustness, even when handling intricate backdrops and multi-scale targets.

4.6. Generalisation Capability of the Model in Jishishan Earthquake

In practical applications, the generalisation capability of a post-earthquake building extraction model is crucial. To evaluate the generalisation performance of the proposed

UEBNet, we conducted tests using UAV imagery (resolution: 0.14 m) of Dahejia Town, captured after the Ms 6.2 Jishishan earthquake in Gansu Province, China, on 18 December 2023. As shown in Figure 12, we collected data from this region to be used for model training and evaluation. Figure 13 shows sample instances from the dataset.

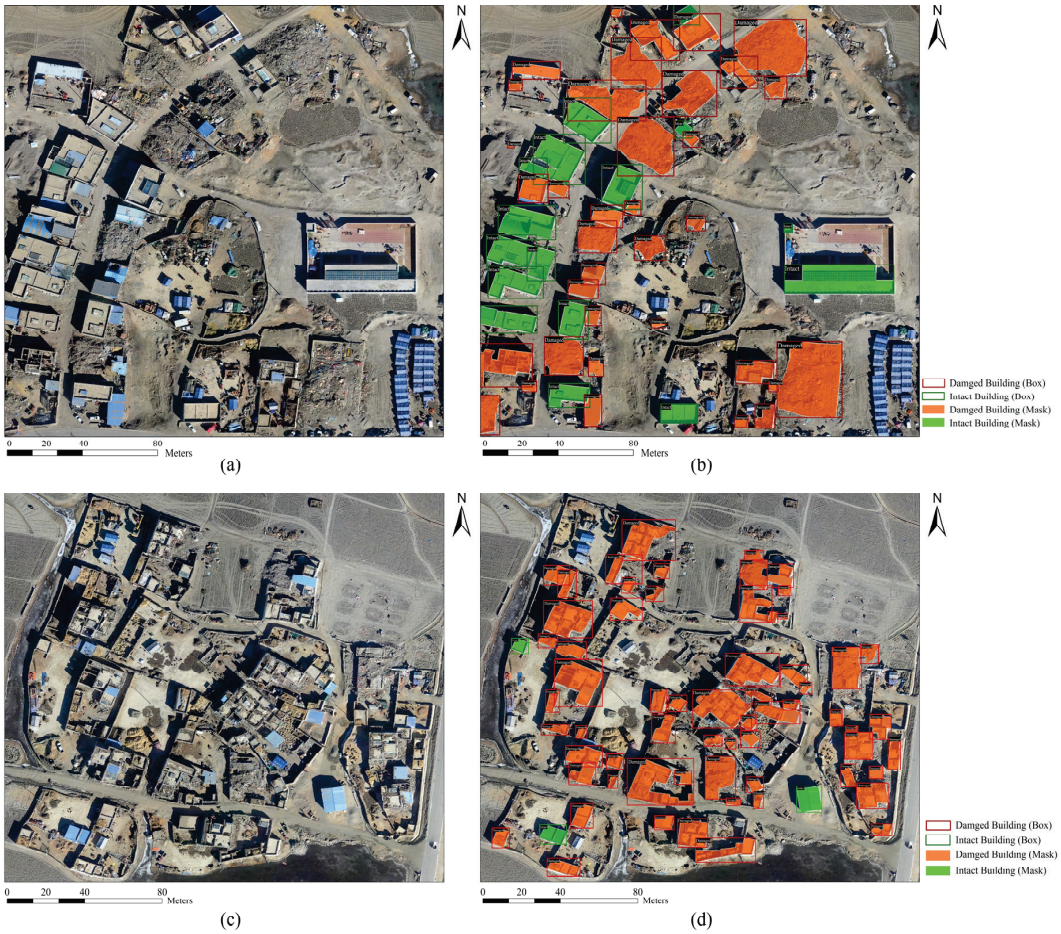


Figure 11. Building extraction results of the UEBNet model for Chajiang Village (**a**,**b**) and Tonglai Village (**c**,**d**), where (**a**,**c**) are the original images, and (**b**,**d**) are the segmentation results.

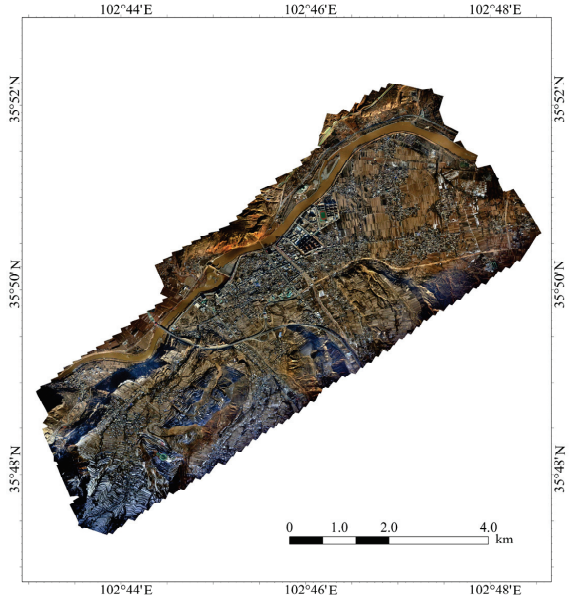


Figure 12. Post-earthquake UAV imagery of the Jishishan earthquake.

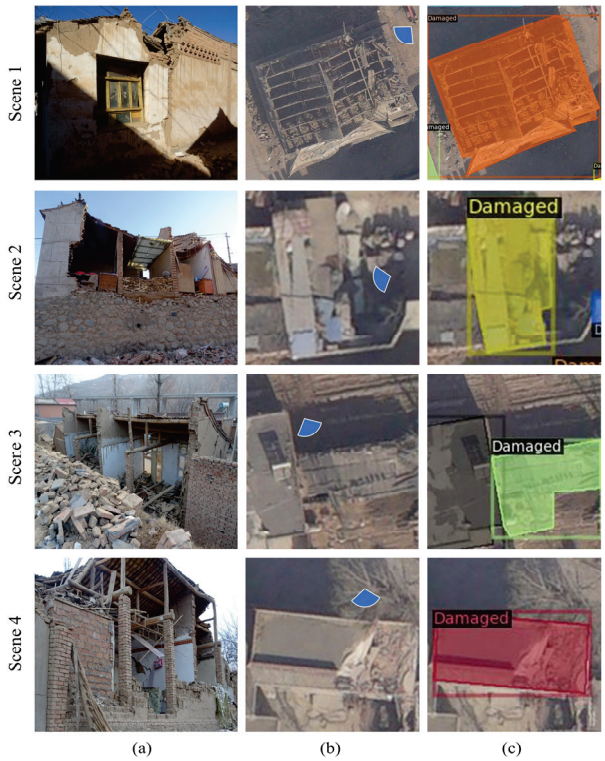


Figure 13. Example of Jishishan earthquake data samples: (**a**) field survey photos; (**b**) UAV images, where the observation location's viewing angle is indicated by blue fan shapes; (**c**) sample labels, with different mask colours representing different instances.

The test results for UEBNet and the SOTA models are presented in Table 4 and Figure 14. The results indicate that UEBNet still achieved the best performance on the Jishishan earthquake data, with an OA of 95.33%. As depicted in Figure 14, UEBNet can accurately locate buildings and capture their irregular boundaries, performing well even in multi-target scenarios, which demonstrates its strong generalisation capability. To further assess UEBNet's practical application performance, we conducted inference tests on parts of two villages in Jishishan County, with the results shown in Figure 15. UEBNet successfully detected and precisely segmented the various types of buildings within these areas, and their boundaries showed a high degree of conformity to the actual outlines. The classification results were consistent with visual interpretation, exhibiting low rates of misclassification and omission. These results fully confirm the generalisation capability of the UEBNet model across different earthquake events.

Table 4. Quantitative results for all models of instance masks. Best results are shown in boldface.

Model	mAP^{box} (%)	$AP^{box}_{0.5}$ (%)	$AP^{box}_{0.75}$ (%)	mAP^{seg} (%)	$AP^{seg}_{0.5}$ (%)	$AP^{seg}_{0.75}$ (%)	F1 (%)	OA (%)	FPS (img/s)
Mask R-CNN	32.84	56.80	33.78	31.83	55.63	32.49	74.59	92.30	23.87
DCT-Mask	33.80	54.91	35.57	31.12	54.88	30.69	75.03	93.61	23.32
YOLOv8-Ins-s	34.10	55.37	34.75	33.07	54.92	31.67	75.58	93.72	**30.26**
PointRend	34.37	58.79	33.71	34.72	57.02	31.97	76.26	94.18	18.32
YOLACT	35.44	59.30	36.92	34.51	56.66	33.75	76.01	93.89	19.20
SOLOv2	36.98	58.61	36.23	36.11	56.75	35.02	77.38	94.46	20.75
PatchDCT	38.09	56.64	37.93	37.57	55.84	36.60	78.12	94.98	14.89
Mask2Former	38.40	57.83	**38.91**	38.21	57.10	37.20	78.45	95.15	14.20
UEBNet (Ours)	**39.32**	**61.28**	38.64	**39.12**	**60.42**	**38.63**	**79.27**	**95.33**	12.57

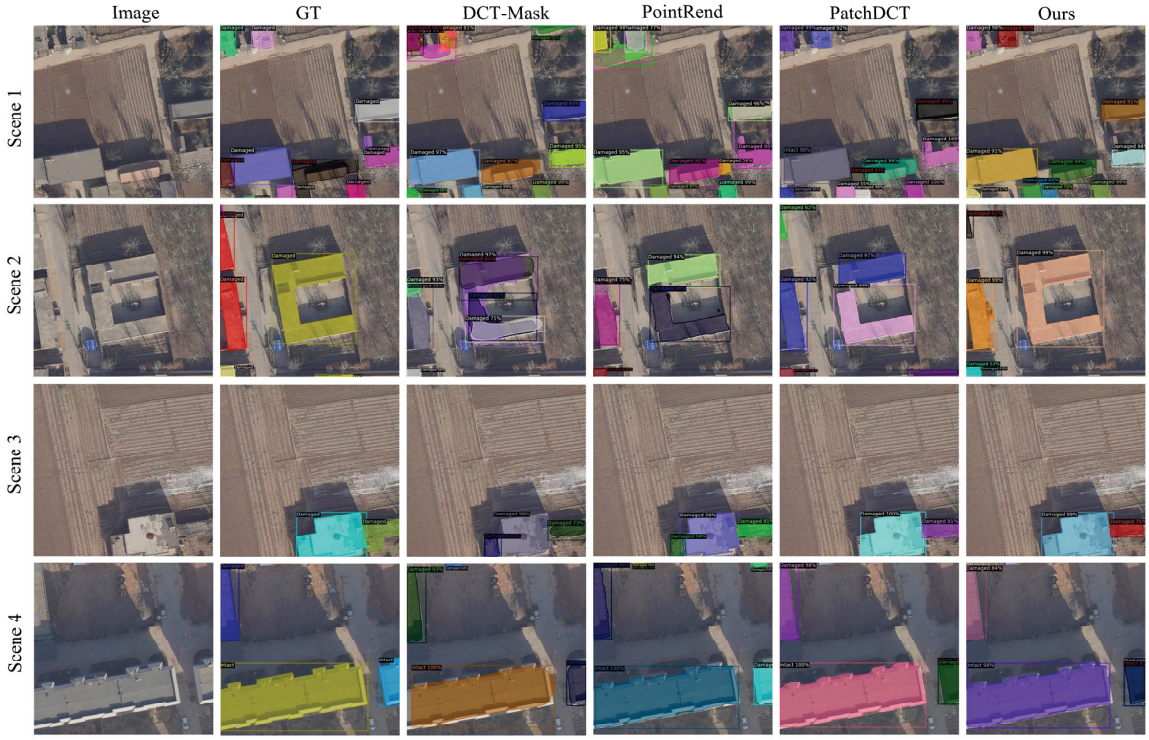


Figure 14. Instance segmentation results of different models on the post-earthquake building dataset from the Jishishan earthquake.

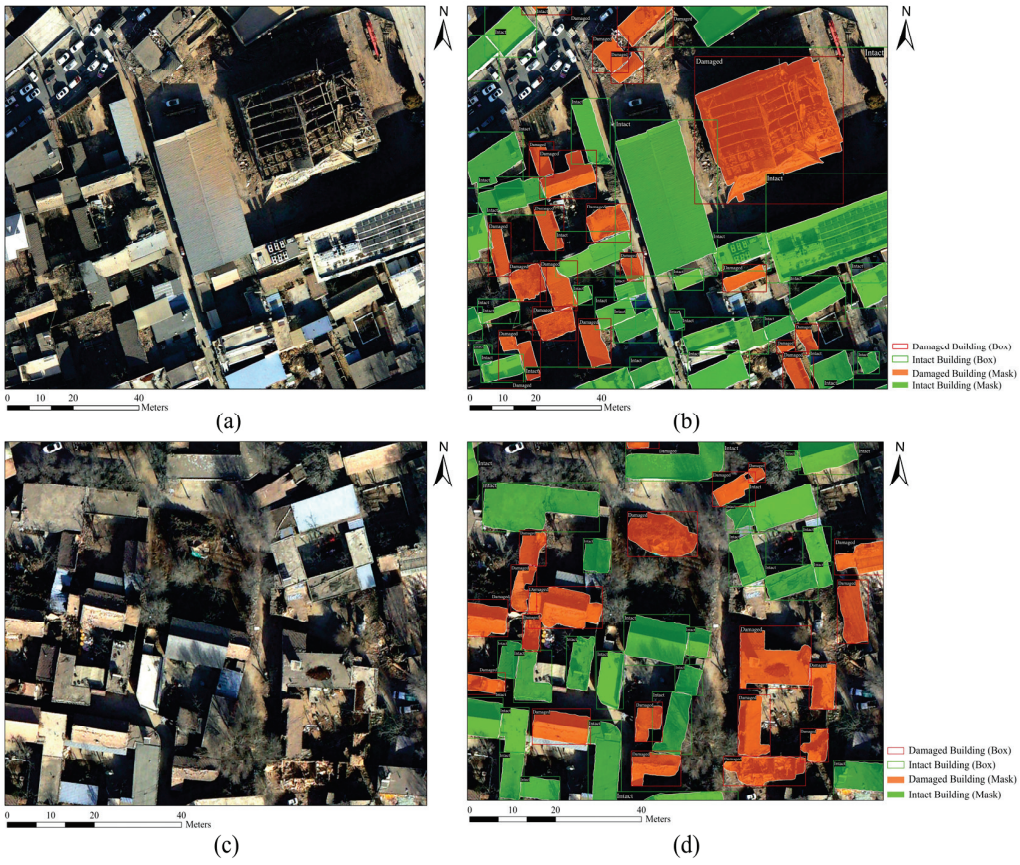


Figure 15. Building extraction results of the UEBNet model for Dahe Village (**a**,**b**) and Zhoujia Village (**c**,**d**) in Jishishan County, where (**a**,**c**) are the original images, and (**b**,**d**) are the segmentation results.

We observed that UEBNet's metrics for the Jishishan earthquake were lower than in the Tingri earthquake scenario. For example, the mAP^{seg} for the Jishishan earthquake

in Table 4 was 39.12%, lower than the 42.26% for the Tingri earthquake in Table 2. This trend was also present in the other models. We believe this performance difference stems primarily from three factors: Firstly, there was a discrepancy in the resolution of the UAV imagery collected from the two earthquakes, with the Tingri imagery having a higher resolution (5 cm). Secondly, due to weather conditions, there was light haze in the Jishishan earthquake, which reduced UAV image clarity. Thirdly, structures in Jishishan County are predominantly timber-and-earth or brick-and-concrete constructions, with most village dwellings featuring interconnected roofs and dense housing layouts. These exhibit localised cracking or partial collapse. In contrast, buildings in Tingri County are chiefly timber-and-earth or earth-and-stone structures, with unconnected dwellings. Most display either complete overall collapse or partial collapse.

4.7. Hyperparameter Sensitivity Analysis

To validate the model's robustness to hyperparameter selection, we conducted a comprehensive parameter sensitivity analysis. We focused primarily on three key hyperparameters within the loss function: λ_0, which controls the weighting of coarse mask regression loss; λ_1, which controls the weighting of fine patch loss; and γ, the focal parameter in Focal Loss used to address class imbalance. Given that the UEBNet model employs a coarse-to-fine optimisation process, where global structure prediction and local detail refinement are equally important, we set $\lambda_0 = \lambda_1$, hereafter referred to as λ. Using the Tingri seismic dataset and fixing all other training settings, we systematically adjusted these parameters, evaluating performance primarily via mAP^{box} and mAP^{seg} metrics. The results are presented in Table 5.

Table 5. Results of hyperparameter sensitivity analysis. Default configurations are indicated with an underscore. Best results are shown in boldface.

$\gamma \backslash \lambda$	mAP^{box} (%)				mAP^{seg} (%)			
	0.25	**0.50**	**0.75**	**1.00**	**0.25**	**0.50**	**0.75**	**1.00**
1.00	39.68	40.38	37.16	36.91	39.24	40.13	37.30	36.85
2.00	39.30	**41.94**	41.56	39.28	38.87	**42.26**	40.93	37.77
3.00	38.95	40.92	37.98	36.58	38.47	40.17	37.05	36.05
4.00	38.72	40.07	37.36	36.02	38.09	39.39	36.81	35.57
5.00	36.23	37.01	36.49	35.67	35.36	37.24	36.03	35.20

The results indicate that the model achieves peak performance under default configurations (λ = 0.50, γ = 2.00). Specifically, when γ = 2.00, performance peaks at λ = 0.50. As λ shifts towards either 0.25 or 1.00, performance declines, demonstrating the critical importance of assigning balanced weights to the coarse mask loss and fine patch loss. Even when γ deviates from 2.00, λ = 0.50 remains the optimal choice, demonstrating the robustness of this value.

When λ = 0.50, performance peaks at γ = 2.00. As γ decreases or increases, both metrics exhibit a downward trend, indicating that a moderate focus factor is crucial for handling class imbalance. Excessively large or small values of γ diminish performance. When λ deviates from 0.50, γ = 2.00 remains the optimal choice, though overall performance declines.

Although performance declines when deviating from the default configuration, accuracy remains above 40% at γ = 2.00 when λ is set to 0.50 and 0.75, demonstrating a degree of robustness. However, performance degradation is severe in extreme parameter regions.

Sensitivity analysis indicates that UEBNet exhibits notable sensitivity to hyperparameters λ and γ, with the default configuration (λ = 0.50, γ = 2.00) representing the optimal

choice. This analysis not only validates our parameter selection but also provides crucial tuning references for researchers applying this method to similar tasks.

5. Conclusions

In this study, we focused on the highly challenging task of instance segmentation for damaged buildings in post-earthquake UAV remote sensing imagery. To address the core difficulties of complex backgrounds, variable target scales, and irregular boundaries in post-disaster scenes, we designed and implemented a novel high-performance segmentation network model, UEBNet. By synergistically optimizing the DSCBAM module, the MFT module, and the AMSI module, the model systematically improves segmentation accuracy. We verified the effectiveness and superiority of UEBNet through extensive experiments. It achieved SOTA performance across all key evaluation metrics, demonstrating exceptional capability, particularly when processing partially damaged and irregularly shaped buildings. UEBNet not only provides an efficient and precise technical tool for post-earthquake emergency response and damage assessment but also offers new design concepts and a robust baseline model for addressing similar complex visual tasks.

Despite its numerous advantages, UEBNet still faces challenges in practical post-earthquake emergency response. Its two-stage instance segmentation architecture requires significant computational resources, which to some extent limits its rapid deployment and real-time processing capabilities in emergency scenarios.

Future work will investigate model simplification and compression to enhance efficiency while maintaining accuracy. Furthermore, considering the limitations of remote sensing observations, we will aim to increase the precision of remote sensing-based seismic damage instance segmentation by integrating on-site investigations and seismological knowledge.

Author Contributions: Conceptualization, Z.G., S.W. and K.Y.; methodology, Z.G. and S.W.; software, Z.G.; validation, Z.G.; formal analysis, Z.G. and S.W.; investigation, Z.G.; resources, S.W. and X.Z.; data curation, Z.G.; writing—original draft preparation, Z.G.; writing—review and editing, S.W., K.Y. and Y.W.; visualization, Z.G.; supervision, S.W.; funding acquisition, S.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Fundamental Research Funds of the Institute of Earthquake Forecasting, China Earthquake Administration, Grant CEAIEF20230202; the National Natural Science Foundation of China, Grant 42271090; and the Fundamental Research Funds of the Institute of Earthquake Forecasting, China Earthquake Administration, Grant CEAIEF20240508 and CEAIEF20240405.

Data Availability Statement: Dataset available on request from the authors.

Acknowledgments: The authors would like to thank the anonymous reviewers for their very competent comments and helpful suggestions.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Li, Q.Y.; Mou, L.C.; Sun, Y.; Hua, Y.S.; Shi, Y.L.; Zhu, X.X. A Review of Building Extraction From Remote Sensing Imagery: Geometrical Structures and Semantic Attributes. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 4702315. [CrossRef]
2. Reuland, Y.; Lestuzzi, P.; Smith, I.F.C. A model-based data-interpretation framework for post-earthquake building assessment with scarce measurement data. *Soil. Dyn. Earthq. Eng.* **2019**, *116*, 253–263. [CrossRef]
3. Valentijn, T.; Margutti, J.; van den Homberg, M.; Laaksonen, J. Multi-Hazard and Spatial Transferability of a CNN for Automated Building Damage Assessment. *Remote Sens.* **2020**, *12*, 2839. [CrossRef]
4. An, L.; Zhang, J.; Ricardo, M.; Zhang, L. A review and prospective research of earthquake damage assessment and remote sensing. *Natl. Remote Sens. Bull.* **2024**, *28*, 860–884. [CrossRef]

5. Fan, Y.; Wu, W.; Wang, W.; Liu, M.; Win, Q. Research progress of disaster remote sensing in China. *Natl. Remote Sens. Bull.* **2021**, *20*, 1170–1184. [CrossRef]
6. Ge, J.Y.; Tang, H.; Yang, N.S.; Hu, Y.J. Rapid identification of damaged buildings using incremental learning with transferred data from historical natural disaster cases. *Isprs J. Photogramm. Remote Sens.* **2023**, *195*, 105–128. [CrossRef]
7. Jhan, J.P.; Kerle, N.; Rau, J.Y. Integrating UAV and Ground Panoramic Images for Point Cloud Analysis of Damaged Building. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 6500805. [CrossRef]
8. Nedjati, A.; Vizvari, B.; Izbirak, G. Post-earthquake response by small UAV helicopters. *Nat. Hazards* **2016**, *80*, 1669–1688. [CrossRef]
9. Wang, J.Q.; Guo, H.N.; Su, X.; Zheng, L.; Yuan, Q.Q. PCDASNet: Position-Constrained Differential Attention Siamese Network for Building Damage Assessment. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5622318. [CrossRef]
10. Xie, Y.K.; Feng, D.J.; Chen, H.Y.; Liu, Z.C.; Mao, W.F.; Zhu, J.; Hu, Y.; Baik, S.W. Damaged Building Detection From Post-Earthquake Remote Sensing Imagery Considering Heterogeneity Characteristics. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 4708417. [CrossRef]
11. Zhang, R.; Li, H.; Duan, K.; You, S.; Liu, K.; Wang, F.; Hu, Y. Automatic detection of earthquake-damaged buildings by integrating UAV oblique photography and infrared thermal imaging. *Remote Sens.* **2020**, *12*, 2621. [CrossRef]
12. Yu, B.; Sun, Y.; Hu, J.; Chen, F.; Wang, L. Post-disaster building damage assessment based on gated adaptive multi-scale spatial-frequency fusion network. *Int. J. Appl. Earth Obs. Geoinf.* **2025**, *141*, 104629. [CrossRef]
13. Hui, J.; Du, M.K.; Ye, X.; Qin, Q.M.; Sui, J. Effective Building Extraction From High-Resolution Remote Sensing Images With Multitask Driven Deep Neural Network. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 786–790. [CrossRef]
14. Li, Y.S.; Li, X.W.; Zhang, Y.J.; Peng, D.F.; Bruzzone, L. Cost-efficient information extraction from massive remote sensing data: When weakly supervised deep learning meets remote sensing big data. *Int. J. Appl. Earth Obs. Geoinf.* **2023**, *120*, 103345. [CrossRef]
15. Mo, S.Y.; Shi, Y.F.; Yuan, Q.; Li, M.Y. A Survey of Deep Learning Road Extraction Algorithms Using High-Resolution Remote Sensing Images. *Sensors* **2024**, *24*, 1708. [CrossRef] [PubMed]
16. He, K.M.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 386–397. [CrossRef] [PubMed]
17. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Munich, Germany, 5–9 October 2025; pp. 234–241. [CrossRef]
18. Cheng, B.W.; Misra, I.; Schwing, A.G.; Kirillov, A.; Girdhar, R. Masked-attention Mask Transformer for Universal Image Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 1280–1289. [CrossRef]
19. Duarte, D.; Nex, F.; Kerle, N.; Vosselman, G. Multi-Resolution Feature Fusion for Image Classification of Building Damages with Convolutional Neural Networks. *Remote Sens.* **2018**, *10*, 1636. [CrossRef]
20. Xiong, C.; Li, Q.; Lu, X. Automated regional seismic damage assessment of buildings using an unmanned aerial vehicle and a convolutional neural network. *Autom. Constr.* **2020**, *109*, 102994. [CrossRef]
21. Khankeshizadeh, E.; Mohammadzadeh, A.; Arefi, H.; Mohsenifar, A.; Pirasteh, S.; Fan, E.; Li, H.X.; Li, J.A.T. A Novel Weighted Ensemble Transferred U-Net Based Model (WETUM) for Postearthquake Building Damage Assessment From UAV Data: A Comparison of Deep Learning- and Machine Learning-Based Approaches. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 4701317. [CrossRef]
22. Cao, X.; Zou, H.; Li, J.; Ying, X.; He, S. OBBInst: Remote sensing instance segmentation with oriented bounding box supervision. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *128*, 103717. [CrossRef]
23. Gomroki, M.; Hasanlou, M.; Chanussot, J.; Hong, D.F. EMYNet-BDD: EfficientViTB Meets Yolov8 in the Encoder-Decoder Architecture for Building Damage Detection Using Postevent Remote Sensing Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 13120–13134. [CrossRef]
24. Zhong, S.; Wang, L. Review of Research on Object Detection in UAV Aerial Images. *Laser Optoelectron. Progress.* **2025**, *62*, 1000005. [CrossRef]
25. Yu, K.; Wang, S.; Wang, Y.; Gu, Z.; Wang, Y. DBA-RTMDet: A High-Precision and Real-Time Instance Segmentation Method for Identification of Damaged Buildings in Post-Earthquake UAV Imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 19577–19593. [CrossRef]
26. Zou, R.; Liu, J.; Pan, H.; Tang, D.; Zhou, R. An improved instance segmentation method for fast assessment of damaged buildings based on post-earthquake uav images. *Sensors* **2024**, *24*, 4371. [CrossRef] [PubMed]
27. Zhan, Y.H.; Liu, W.; Maruyama, Y. Damaged Building Extraction Using Modified Mask R-CNN Model Using Post-Event Aerial Images of the 2016 Kumamoto Earthquake. *Remote Sens.* **2022**, *14*, 1002. [CrossRef]

28. Shen, X.; Yang, J.; Wei, C.; Deng, B.; Huang, J.; Hua, X.-S.; Cheng, X.; Liang, K. Dct-mask: Discrete cosine transform mask representation for instance segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 8720–8729. [CrossRef]
29. Wen, Q.; Yang, J.; Yang, X.; Liang, K. Patchdct: Patch refinement for high quality instance segmentation. In Proceedings of the Eleventh International Conference on Learning Representations, Kigali, Rwanda, 1–5 May 2023. [CrossRef]
30. Shang, Q.; Zhou, Z. Rapid report of seismic damage and consequence analysis in the 2025 M 6.8 Dingri earthquake. *Earthq. Res. Adv.* **2025**, 100394. [CrossRef]
31. Xu, X.; Wang, S.; Cheng, J.; Wu, X. Shaking the Tibetan Plateau: Insights from the Mw 7.1 Dingri earthquake and its implications for active fault mapping and disaster mitigation. *npj Nat. Hazards* **2025**, *2*, 16. [CrossRef]
32. Can, Z.; Hongmei, G.; Dongming, W.; Yuping, Y.; Zhen, Z.; Ying, Z.; Zonghang, H. Rapid disaster assessment of a 6.8-magnitude earthquake in Dingri, Tibet. *Earthq. Res. Adv.* **2025**, 100377. [CrossRef]
33. Vabalas, A.; Gowen, E.; Poliakoff, E.; Casson, A.J. Machine learning algorithm validation with a limited sample size. *PLoS ONE* **2019**, *14*, e0224365. [CrossRef] [PubMed]
34. Zhou, Z.-H. *Machine Learning*; Springer Nature: Berlin/Heidelberg, Germany, 2021.
35. Woo, S.H.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the 15th European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19. [CrossRef]
36. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258. [CrossRef]
37. Yu, F.; Koltun, V. Multi-Scale Context Aggregation by Dilated Convolutions. In Proceedings of the ICLR, San Juan, Puerto Rico, 2–4 May 2016. [CrossRef]
38. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Honolulu, HI, USA, 21–26 July 2017; pp. 2980–2988. [CrossRef]
39. Tian, Z.; Shen, C.H.; Chen, H.; He, T. FCOS: A Simple and Strong Anchor-Free Object Detector. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 1922–1933. [CrossRef]
40. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874. [CrossRef]
41. Varghese, R.; Sambath, M. YOLOv8: A Novel Object Detection Algorithm with Enhanced Performance and Robustness. In Proceedings of the 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS), Chennai, India, 18–19 April 2024; pp. 1–6. [CrossRef]
42. Kirillov, A.; Wu, Y.X.; He, K.M.; Girshick, R. PointRend: Image Segmentation as Rendering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 9796–9805. [CrossRef]
43. Bolya, D.; Zhou, C.; Xiao, F.Y.; Lee, Y.J. YOLACT Real-time Instance Segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9156–9165. [CrossRef]
44. Wang, X.; Zhang, R.; Kong, T.; Li, L.; Shen, C. SOLOv2: Dynamic and Fast Instance Segmentation. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 17721–17732. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

remote sensing

MDPI

Article

Probabilistic Forecasting Model for Tropical Cyclone Intensity Based on Diffusion Model

Jingjia Luo [1,2], Peng Yang [1,2] and Fan Meng [1,2,*]

[1] State Key Laboratory of Climate System Prediction and Risk Management (CPRM), Nanjing University of Information Science and Technology, Nanjing 210044, China
[2] School of Artificial Intelligence, Nanjing University of Information Science and Technology, Nanjing 210044, China
* Correspondence: meng@nuist.edu.cn

Highlights:

What are the main findings?

- This study proposes a novel conditional diffusion model (TCDM), representing the first probabilistic generative framework for tropical cyclone (TC) intensity forecasting capable of integrating multimodal data to generate the full future intensity probability distribution.
- Compared with mainstream baseline models, TCDM demonstrates a substantial improvement in forecasting rapid intensification (RI) events, achieving the highest hit rate (30.7%) and precision (43.2%) and the lowest false alarm rate (2.7%) in the testing experiments.

What is the implication of the main finding?

- The TCDM framework is capable of generating high-quality and highly reliable probabilistic forecasts, providing more effective decision support for meteorological agencies in tropical cyclone risk assessment and disaster preparedness.
- This study demonstrates the strong potential of diffusion models in simulating and forecasting the uncertainty of extreme weather events, opening new avenues for developing more intelligent and reliable meteorological disaster forecasting systems.

Abstract

Reliable forecasting of tropical cyclone (TC) intensity—particularly rapid intensification (RI) events—remains a major challenge in meteorology, largely due to the inherent difficulty of accurately quantifying predictive uncertainty. Traditional numerical approaches are computationally expensive, while statistical models often fail to capture the highly nonlinear relationships involved. Mainstream machine learning models typically provide only deterministic point forecasts and lack the ability to represent uncertainty. To address this limitation, we propose Tropical Cyclone Diffusion Model (TCDM), the first conditional diffusion-based probabilistic forecasting framework for TC intensity. TCDM integrates multimodal meteorological data, including satellite imagery, re-analysis fields, and environmental predictors, to directly generate the full probability distribution of future intensities. Experimental results show that TCDM not only achieves highly competitive deterministic accuracy (low MAE and RMSE; high R^2), but also delivers high-quality probabilistic forecasts (low CRPS; high PICP). Moreover, it substantially improves RI detection by achieving higher hit rates with fewer false alarms. Compared with traditional ensemble-based methods, TCDM provides a more efficient and flexible approach to probabilistic forecasting, offering valuable support for TC risk assessment and disaster preparedness.

Keywords: tropical cyclone intensity forecasting; rapid intensification (RI); probabilistic prediction; TCDM; diffusion model; multimodal data

1. Introduction

Tropical cyclones represent one of the most destructive natural hazards worldwide. They pose severe threats to human life, infrastructure, and socio-economic systems along coastal regions through extreme winds, storm surges, and heavy rainfall. Over the past five decades, TCs have been responsible for the largest economic losses among all weather-related hazards [1], with catastrophic events such as Typhoon Haiyan (2013) and Hurricane Katrina (2005) vividly demonstrating their immense destructive potential [2,3]. Accurate forecasting of TC intensity, particularly rapid intensification (RI) events characterized by abrupt and substantial strengthening [4,5], is therefore critical for disaster preparedness, early warning, and risk mitigation. Despite significant advancements in numerical weather prediction (NWP) models [6] and statistical forecasting methods [7,8], predicting TC intensity—especially RI events and medium- to long-term trends—remains a challenging problem. Dynamical models are heavily dependent on high-resolution simulations and precise initial conditions, yet they often fail to fully capture small-scale physical processes that are crucial for accurate representation of TC structure and evolution [9]. In contrast, statistical models, such as the Statistical Hurricane Intensity Prediction Scheme (SHIPS) [10,11], offer computational efficiency but are limited in representing the complex nonlinear interactions between TC dynamics and the surrounding environmental conditions [12]. These intrinsic limitations contribute to substantial forecast uncertainty [13], which in turn constrains the practical utility of operational predictions. In recent years, numerical weather prediction (NWP) models have achieved remarkable progress in rapid intensification (RI) forecasting. This advancement has been driven by continuous improvements in computational resources and a deeper understanding of tropical cyclone (TC) dynamics and physical processes. For instance, the new generation of high-resolution dynamical models—such as the Hurricane Analysis and Forecast System (HAFS) and the updated version of the Hurricane Weather Research and Forecasting model (HWRF)—have demonstrated stronger competitiveness in RI forecasting skills by incorporating more advanced data assimilation schemes and physical process parameterizations [14,15]. However, due to the highly nonlinear dynamical characteristics of TC systems and their extreme sensitivity to initial conditions and model physics, NWP models still face great challenges in accurately quantifying the probability and uncertainty of RI events. Moreover, the computational cost of high-resolution ensemble forecasting systems remains extremely expensive.

In recent years, the rapid growth of machine learning (ML), particularly deep learning (DL), has demonstrated substantial potential in geoscientific applications [16,17]. In many complex meteorological tasks—such as El Niño prediction [18] and short-term precipitation forecasting [19]—deep learning models have outperformed conventional methods. Within the context of TC intensity forecasting, deep learning architectures, including convolutional neural networks (CNNs), recurrent neural networks (RNNs), and their variants such as ConvLSTM [20], have been widely adopted [21,22]. These models leverage large-scale spatiotemporal datasets, encompassing both satellite imagery [23] and re-analysis products, to extract meaningful patterns and improve prediction accuracy [24–26]. However, most existing ML approaches remain deterministic, providing point estimates [27,28] and failing to capture the inherent uncertainties in TC intensity evolution. Such uncertainties arise from chaotic atmospheric dynamics, observational noise, and intrinsic model limitations [29].

Probabilistic forecasting, which offers confidence intervals and full probability density estimates [30], is essential for comprehensive risk assessment, informed decision-making, and

adaptive emergency management [31]. Operational centers frequently use ensemble-based approaches to quantify forecast uncertainty [13], generating multiple realizations by perturbing initial conditions or combining different model outputs. Nonetheless, ensemble approaches demand significant computational resources, and limited ensemble sizes may result in incomplete or biased representation of probabilistic outcomes. Alternative statistical techniques, such as Bayesian frameworks or quantile regression [32], provide probabilistic information but often rely on simplified distributional assumptions or extensive prior knowledge, limiting their effectiveness for complex, multimodal TC datasets.

Diffusion models, as an emerging class of generative machine learning techniques, iteratively denoise data through a Markov chain process to learn high-dimensional distributions. Recently, they have achieved remarkable success in modeling complex probabilistic structures and generating realistic, diverse samples [33–36]. Compared with Generative Adversarial Networks (GANs)[37], diffusion models exhibit more stable training and greater diversity in generated samples; relative to Variational Autoencoders (VAEs) [38], they demonstrate a significant advantage in preserving data fidelity. By capturing the intrinsic spatiotemporal patterns present in meteorological data, diffusion models offer a promising avenue for probabilistic TC forecasting, particularly in scenarios characterized by high uncertainty and nonlinear interactions.

In this study, we introduce Tropical Cyclone Diffusion Model, a novel conditional diffusion framework designed specifically for probabilistic TC intensity prediction. The proposed model integrates multimodal meteorological information—including re-analysis fields, satellite observations, and best-track records—to jointly forecast future intensity trajectories and quantify associated uncertainties. The core architecture employs a conditional diffusion-generation process, allowing for effective feature extraction and fusion across heterogeneous data sources. This integration enhances both the robustness and interpretability of the predictions, providing a unified framework for high-fidelity probabilistic forecasting of tropical cyclone intensity.

2. Data and Methodology

2.1. Data

2.1.1. Satellite Imagery Data

The satellite observational data used in this study were obtained from the Tropical Cyclone Integrated Reanalysis (TCIR) dataset [39], which is publicly available through the Intelligent Systems Laboratory at National Taiwan University. This dataset establishes a comprehensive tropical cyclone feature archive by integrating multi-source satellite remote sensing observations. Its core data sources include the GridSat global infrared dataset and the CMORPH precipitation dataset. Specifically, GridSat has continuously archived observations from global geostationary satellites at a 3-h interval since 1981, providing brightness temperature data in the infrared window channel (IR1, 11.2 μm), water vapor channel (WV, 6.7 μm), and visible channel (VIS, 0.6 μm), with a spatial resolution of 0.07° × 0.07°. CMORPH, on the other hand, generates global precipitation rate fields from 2003 to 2017 by combining low-Earth-orbit microwave satellite measurements with geostationary infrared data through a morphing technique, with a spatial resolution of 0.25° × 0.25° and a temporal resolution of 3 h.

For the validation of tropical cyclone tracks and intensities, the best-track data from the Joint Typhoon Warning Center (JTWC) are used for the western North Pacific, while the revised Atlantic Hurricane Database (HURDAT2) provides re-analysis records for the eastern North Pacific and Atlantic basins from 2003 to 2017. By integrating these multi-source datasets, TCIR provides key tropical cyclone parameters, including storm center location, intensity measured as maximum sustained wind speed (knots), storm

size expressed as the quadrant-averaged radius of 35 kt winds (nautical miles, NMI), and minimum sea-level pressure (Figure 1). These parameters form a complete feature vector set that serves as a benchmark for training and validating machine learning models.

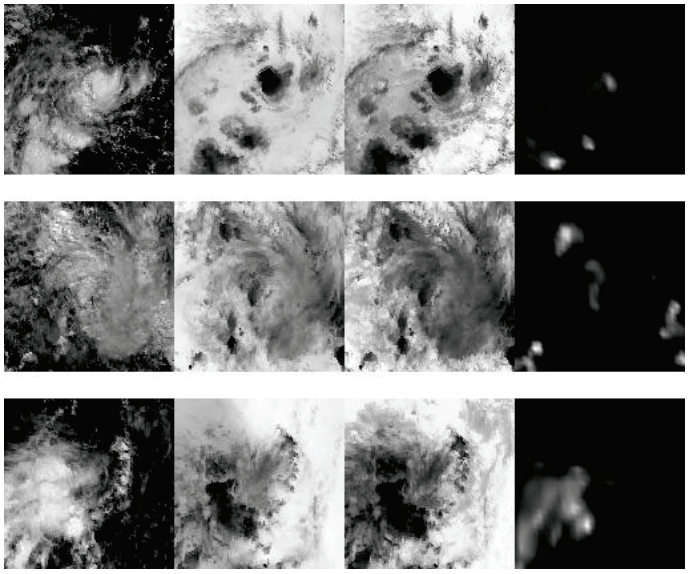


Figure 1. TCRI schematic diagram; from left to right: IR, PMW, VIS, and WV.

2.1.2. SHIPS Development Database

To enhance the model's ability to capture the environmental context of tropical cyclones, this study incorporates the Statistical Hurricane Intensity Prediction Scheme (SHIPS) developmental database, which was jointly developed by the National Hurricane Center (NHC) and Colorado State University. The SHIPS database integrates a broad set of environmental predictors that influence the evolution of tropical cyclone intensity, and it has been widely employed in both operational forecasting and model training. Specifically, it includes key physical variables such as sea surface temperature (SST), vertical wind shear at upper and mid-levels, relative humidity (RH), ocean heat content (OHC), latent heat flux, wind field symmetry, pressure gradients, and geopotential height.

The SHIPS dataset covers historical tropical cyclone processes across multiple ocean basins, with environmental predictors derived from NCEP global re-analysis products (e.g., GFS and CFSR). Its primary advantages lie in the standardized data structure and strong physical interpretability, making it highly suitable for constructing statistical or machine learning models based on environmental fields. In this study, SHIPS samples from 2003 to 2017, temporally aligned with TCIR satellite imagery, were employed. Furthermore, the key variables were standardized to ensure seamless fusion with remote sensing image features within the proposed model.

By introducing SHIPS predictors, the model is able to more comprehensively capture the thermodynamic and dynamic mechanisms underlying cyclone development. This integration effectively enhances the physical consistency and generalization ability of intensity forecasts, while providing critical prior information in rapid intensification scenarios.

2.1.3. Data Preprocessing and Alignment

The raw dataset, which includes both satellite imagery and tabular information, was first processed to remove any observations lacking SHIPS environmental records and to handle NaN or anomalous values, such as truncating visible (VIS) channel intensities to the range [0, 1]. Subsequently, the data were partitioned into training, validation, and test sets according to cyclone ID in temporal order, ensuring that the complete historical sequence of a single cyclone did not leak across datasets and thereby preventing information leakage

and distribution bias. For each tropical cyclone case, a time series was constructed consisting of synchronized satellite images and SHIPS forecast factors, ensuring spatiotemporal consistency and completeness of the input data for reliable model training.

During model training, specific image channels were selected according to the experimental configuration, and all images were standardized to uniform dimensions and value ranges to ensure stable and efficient training. To enhance rotational invariance and generalization capability, multiple image augmentation strategies were employed, including random-angle rotation and physics-based rotations aligned with the vertical wind shear direction (SHTD) at 0°, 90°, 180°, or 270°. These augmentations allow the model to effectively learn the relative spatial relationships between cyclone structures and environmental wind fields while preserving the accuracy of the physical information.

2.2. TCDM Model for Probabilistic Forecasting of Tropical Cyclone Intensity

The TCDM architecture, depicted in Figure 2, is designed to effectively fuse information from different sources and leverage the generative power of diffusion models for probabilistic forecasting. The core components are the feature encoder, the Condition Fusion module, and the conditional diffusion process itself.

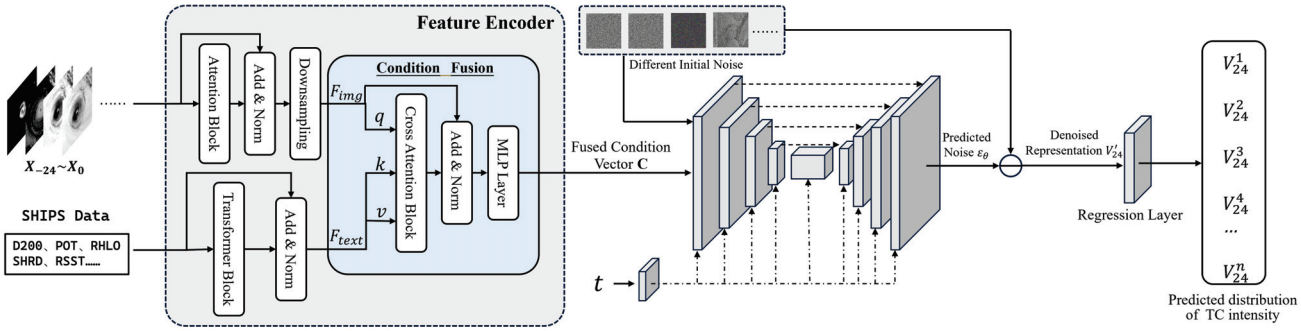


Figure 2. Schematic diagram of the proposed TCDM framework.

2.2.1. Feature Encoder

The feature encoder processes two primary data streams: multi-channel satellite imagery and SHIPS environmental predictors.

For the four-channel satellite images, the model employs a series of attention blocks. To ensure the robustness of feature extraction, each attention block employed in this study consists of a Multi-Head Self-Attention (MHSA) module and a position-wise feed-forward network (FFN). Each module incorporates residual connections and layer normalization to stabilize training and facilitate gradient flow. In total, four such attention blocks are stacked, each utilizing eight attention heads, enabling the model to capture complex spatial dependencies within satellite imagery, such as convective organization and eyewall structures. The extracted features are subsequently downsampled to produce a compact yet information-rich image feature vector, denoted as F_{img}.

For the SHIPS data, a Transformer block is adopted. The self-attention mechanism within the Transformer is particularly effective at capturing dependencies among environmental predictors and identifying the most influential factors. The Transformer block adopted in this study was implemented as a standard Transformer encoder layer. It consists of a Multi-Head Self-Attention module and a feed-forward network, both equipped with residual connections and layer normalization to enhance training stability. Two encoder layers are stacked, each employing eight attention heads and a 512-dimensional feed-forward network. Prior to feeding the SHIPS variable vectors into the encoder, a learnable positional encoding is introduced to help the model capture the sequential dependencies and relative importance among different environmental predictors, such as vertical wind shear (VWS),

sea surface temperature (SST), and relative humidity (RH). This process yields a robust textual feature vector, denoted as F_{text}.

2.2.2. Condition Fusion

The fusion of multimodal information is accomplished through a cross-attention mechanism. Specifically, the image feature vector F_{img} and the textual feature vector F_{text} attend to each other, enabling the model to learn conditional relationships between structural features of TCs and environmental constraints. For example, given a specific vertical wind shear condition derived from SHIPS predictors, the model is able to infer potential evolutions in the cloud organization observed in satellite imagery.

The fused representation is subsequently passed through a multilayer perceptron (MLP), producing the final condition vector C. This vector encapsulates a comprehensive meteorological context and is used to guide the denoising process of the diffusion model, thereby enhancing the accuracy and reliability of probabilistic TC intensity forecasts.

2.2.3. Diffusion Model Framework

Diffusion models [33] are a class of models that use forward and reverse processes to reconstruct target data, aiming to recover the original information from noisy inputs. In our task, the objective is to restore the tropical cyclone intensity after 24 h from noisy data. The core of the diffusion process comprises the following two components:

- Forward diffusion process: In the forward process, the model gradually adds noise to the true intensity information until it becomes pure noise. Assuming the initial true intensity, noise is added at each time step to obtain the noisy intensity, and this process can be represented as

$$q(x_t \mid x_0) = \mathcal{N}(x_t; \sqrt{\bar{a}_t}x_0, (1 - \bar{a}_t)I) \tag{1}$$

  In this context, $\mathcal{N}(\cdot)$ denotes the Gaussian distribution, $\bar{a}_t$ serves as the noise attenuation factor that evolves with the time step t, and I denotes the identity matrix, which characterizes the variance of the noise.

- Reverse process: The reverse denoising process aims to recover the true signal intensity from noise by progressively removing noise through the prediction of the noise component at each step, ultimately resulting in the final intensity estimate. This reverse denoising procedure can be mathematically described by the following formula:

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)I) \tag{2}$$

  Here, $\mu_\theta(x_t, t)$ represents the denoised estimate predicted based on the current noisy state x_t and the time step t, while $\Sigma_\theta(x_t, t)$ denotes the variance of the noise, characterizing the uncertainty in the denoising process.

2.2.4. Conditional Diffusion Mechanism

To enhance the model's adaptability to environmental variations, we incorporate four-channel satellite remote sensing images and SHIPS data as the conditional information C [40]. During the denoising process at each time step, the conditional information C interacts with the current noisy intensity x_t, ensuring that the model can perform intensity predictions under diverse meteorological conditions. The conditional diffusion mechanism of the model can be formulated as follows:

$$p_\theta(x_{t-1} \mid x_t, C) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t, C), \Sigma_\theta(x_t, t)I) \tag{3}$$

Here, c denotes the fused multimodal conditional information, which includes both image and textual features, while $\mu_\theta(x_t, t, C)$ explicitly indicates that the denoising mean is conditioned on C.

2.2.5. Probabilistic Forecast Generation via Diffusion Sampling

A key capability of the proposed *TCDM* framework lies in its ability to generate probabilistic (ensemble) forecasts. This is achieved by exploiting the inherent stochasticity of the reverse diffusion process. To produce an ensemble of N forecasts, the model begins with N distinct initial noise samples, $\{x_T^1, x_T^2, \ldots, x_T^N\}$, independently drawn from the prior distribution $\mathcal{N}(0, I)$.

While keeping the conditional information C—derived from the input satellite imagery and SHIPS environmental predictors—fixed, the model performs the reverse denoising process N times, once for each initial noise sample. This procedure generates N distinct denoised realizations, $\{V_{24}^1, V_{24}^2, \ldots, V_{24}^N\}$, as illustrated in Figure 2. These realizations collectively form the predictive distribution of the 24-h tropical cyclone (TC) intensity. Consequently, this ensemble provides a natural means of quantifying forecast uncertainty.

2.2.6. Loss Function

For the specific and critical forecasting issue of the rapid intensification (RI) of tropical cyclones, the rarity of RI events within the cyclone lifecycle results in a significantly smaller number of RI samples compared to non-RI samples in the training dataset. This constitutes a typical class imbalance problem. Using the standard mean squared error (MSE) as the loss function may cause the model to favor minimizing errors for the majority class (non-RI events) during optimization while insufficiently addressing prediction biases for the minority class (RI events), thereby hindering the model's ability to effectively identify and forecast actual RI occurrences.

To address this challenge, this study employs a weighted mean squared error loss function [4]. Its mechanism involves assigning differentiated weights to samples of different classes, thereby amplifying the contribution of errors from specific sample types during the calculation of the total loss. The reconstructed loss function is

$$w = exp(\frac{\delta V_{max} - 20}{10}) + 1 \tag{4}$$

$$L = E_{t,x_0,\epsilon\sim\mathcal{N}(0,I)}[w \cdot \parallel \epsilon - \epsilon_\theta(x_t, t, C) \parallel^2] \tag{5}$$

Here, $\delta V_{\max}$ denotes the wind speed increment, t is the time step, x_0 represents the true wind speed, ϵ is the ground truth noise sampled from $\mathcal{N}(0, I)$, which was added during the forward diffusion process to generate the noisy wind speed x_t. Moreover, ϵ_θ represents the noise predicted by the denoising network $\epsilon_\theta(x_t, t, C)$, which is trained to approximate ϵ, and C denotes the fused conditional vector.

2.2.7. Training Details

All experiments were conducted using Visual Studio Code (https://code.visualstudio.com/), Python 3, and TensorFlow (https://www.tensorflow.org/). We employed a batch size of 50 and a learning rate of 10^{-4}. To evaluate the model's generalization ability and mitigate overfitting, the dataset was split into training, validation, and test sets corresponding to the periods 2003–2014, 2015–2016, and 2017, respectively. The data from 2003 to 2014 were used as the training set, those from 2015 to 2016 as the validation set, and those from 2017 as the test set. The models used in the experiments were trained on a single NVIDIA A800 GPU for a total of 300 epochs. Regarding the specific parameters of the diffusion model, we set the total number of time steps to $T = 1000$. A linear noise schedule was adopted, where β

increases linearly from $\beta_1 = 10^{-4}$ to $\beta_T = 0.02$. During the sampling phase, the standard DDPM sampling algorithm was employed to generate predictive samples from pure Gaussian noise x_T through an iterative denoising process.

2.3. Evaluation Methods

To quantitatively assess the performance of the proposed TCDM model, we employed both deterministic and probabilistic metrics. Deterministic accuracy was evaluated using the mean absolute error (MAE), root mean square error (RMSE), and the coefficient of determination (R^2), all computed based on the predicted wind speed mean ($\hat{\mu}$) of the 24-h forecast. MAE measures the average magnitude of the prediction errors, RMSE gives higher weight to larger errors, and R^2 quantifies the proportion of variance in the observed values explained by the predicted mean. The formulas are defined as follows:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\hat{\mu}_i - y_i|, \tag{6}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{\mu}_i - y_i)^2}, \tag{7}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(\hat{\mu}_i - y_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}, \tag{8}$$

where y_i denotes the observed wind speed, $\hat{\mu}_i$ denotes the predicted mean, $\bar{y}$ is the mean of observations, and N is the number of samples.

Probabilistic performance was assessed using the continuous ranked probability score (CRPS) and the prediction interval coverage probability (PICP). CRPS measures the difference between the predicted cumulative distribution and the observed value, providing a sharpness and calibration evaluation of the probabilistic forecast. Normally distributed predictions are computed as

$$\mathrm{CRPS}(\hat{\mu}_i, \hat{\sigma}_i; y_i) = \hat{\sigma}_i\left[z_i\big(2\Phi(z_i) - 1\big) + 2\phi(z_i) - \frac{1}{\sqrt{\pi}}\right], \quad z_i = \frac{y_i - \hat{\mu}_i}{\hat{\sigma}_i}, \tag{9}$$

where $\hat{\sigma}_i$ is the predicted standard deviation, and Φ and ϕ are the cumulative distribution function and probability density function of the standard normal distribution, respectively.

PICP evaluates the proportion of observations falling within a nominal $(1-\alpha)\times 100\%$ prediction interval:

$$\mathrm{PICP}_\alpha = \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\big(y_i \in [\hat{\mu}_i - z_{1-\alpha/2}\hat{\sigma}_i, \hat{\mu}_i + z_{1-\alpha/2}\hat{\sigma}_i]\big), \tag{10}$$

where α denotes the significance level, and $z_{1-\alpha/2}$ is the corresponding quantile of the standard normal distribution. In this study, we set $\alpha = 0.05$, corresponding to a 95% prediction interval ($\mathrm{PICP}_{95\%}$). A PICP value closer to 0.95 indicates more reliable interval coverage.

3. Results

3.1. Model Performance

To systematically quantify the contribution of distinct input features, we performed an extensive evaluation of multiple channel configurations, encompassing IR, VIS, PMW, and WV imagery. The comprehensive results of this comparison are presented in Table 1. Among all tested configurations, the integration of IR, PMW, and WV consistently demonstrated the highest performance across both deterministic metrics and probabilistic forecast quality indicators. This enhanced performance can be largely attributed to the complemen-

tary physical information inherently captured by these channels. Specifically, IR imagery provides continuous global coverage of convective cloud-top structures and thermal patterns, allowing for the monitoring of large-scale convective organization and intensity variations. PMW observations, on the other hand, can penetrate cloud layers to reveal precipitation intensity, inner-core dynamics, and latent heat release processes, which are critical for capturing the structural evolution of tropical cyclones. Meanwhile, WV imagery characterizes the surrounding atmospheric moisture environment, offering valuable context for understanding cyclonic intensification and potential rapid development. In contrast, VIS imagery, despite its high spatial resolution and ability to capture fine-scale structural details during daylight, is constrained by diurnal availability and shows reduced sensitivity to deep convective processes, which can introduce additional uncertainty when integrated into predictive models. Collectively, these findings highlight the importance of leveraging complementary multi-channel information to improve both the accuracy and reliability of tropical cyclone intensity forecasts.

Table 1. Performance of the model using different input feature combinations.

Combination	MAE	RMSE	R^2	CRPS	PICP
IR	15.20	19.50	0.60	14.50	0.60
VIS	16.50	20.50	0.50	16.00	0.55
PMW	13.90	17.50	0.66	12.60	0.68
WV	14.80	18.00	0.57	14.80	0.61
IR + VIS	13.50	17.20	0.67	13.50	0.66
IR + PMW	11.80	15.00	0.75	8.80	0.82
IR + WV	13.20	16.70	0.68	11.90	0.70
VIS + PMW	14.20	17.80	0.64	12.20	0.69
VIS + WV	15.20	18.80	0.59	13.60	0.64
PMW + WV	13.60	17.20	0.70	9.80	0.75
IR + VIS + PMW	11.10	13.70	0.77	7.90	0.84
IR + VIS + WV	12.90	16.00	0.68	11.00	0.72
IR + PMW + WV	10.20	12.90	0.77	**7.10**	0.91
VIS + PMW + WV	11.20	14.20	0.72	8.50	0.86
IR + VIS + PMW + WV	10.50	12.80	0.76	7.40	0.90
IR + PMW + WV + SHIPS	**10.04**	**12.73**	**0.78**	7.17	**0.93**

Further incorporation of environmental predictors derived from the SHIPS (Statistical Hurricane Intensity Prediction Scheme) dataset led to systematic and consistent improvements across all evaluation metrics. Specifically, deterministic errors, such as MAE and RMSE, were significantly reduced, while the predictive distributions exhibited enhanced sharpness and improved calibration, as evidenced by PICP values closely approaching the nominal confidence levels. These improvements can be largely attributed to the complementary nature of the SHIPS variables, which include key environmental factors such as vertical wind shear, sea surface temperature (SST), mid-level humidity, and ocean heat content. Collectively, these variables provide an essential large-scale environmental context that is often inadequately captured by satellite observations alone, thereby enabling the model to better represent the physical mechanisms governing tropical cyclone evolution. The inclusion of these environmental constraints is particularly critical in RI events, where favorable or adverse large-scale conditions can exert decisive influence on storm development. For instance, low vertical wind shear and elevated ocean heat content can create conditions conducive to rapid intensification, whereas dry, mid-level air or high shear can suppress cyclone strengthening. By explicitly incorporating these environmental factors, the model is able to adjust its probabilistic forecasts in accordance with the surrounding atmospheric and oceanic conditions, leading to more physically consistent predictions.

Overall, these results underscore the importance of multimodal data fusion, whereby high-resolution satellite imagery is integrated with physically meaningful environmental predictors to enhance forecast skill. This approach not only improves deterministic and probabilistic performance but also increases the robustness and reliability of tropical cyclone intensity predictions, highlighting the substantial operational value of combining multiple complementary observational sources in contemporary forecasting frameworks.

In addition, we systematically investigated the influence of ensemble member size on probabilistic forecasting skill, evaluating performance using MAE, RMSE, CRPS, and PICP metrics. As shown in Figure 3, increasing the number of ensemble members progressively enhanced model skill, with improvements gradually plateauing when the ensemble reached approximately 120 members. At this configuration, deterministic errors were minimized, the CRPS achieved its optimal value, and PICP values most closely aligned with the nominal confidence levels, indicating well-calibrated predictive distributions. These results suggest that a moderate ensemble size provides an effective balance between adequately capturing forecast uncertainty and avoiding redundancy-induced noise that may degrade probabilistic accuracy. Moreover, maintaining this balance is crucial for computational efficiency, particularly in operational forecasting systems where timely predictions are essential. Overall, the findings highlight that careful selection of ensemble size is a key factor in achieving both reliable and computationally feasible probabilistic tropical cyclone intensity forecasts.

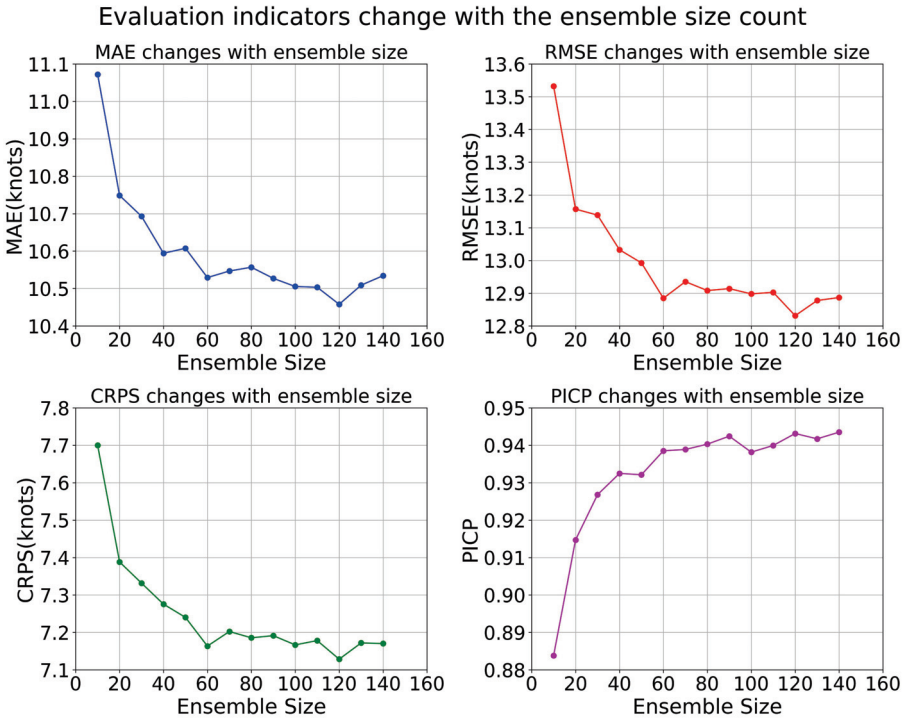


Figure 3. Performance of the model using different ensemble sizes.

The model's predictive performance was comprehensively evaluated through multiple complementary analyses (Figure 4). The predicted–observed scatter closely follows the 1:1 line, with an overall coefficient of determination of $R^2 = 0.78$, indicating robust explanatory power and stable performance across the test set, while residuals remain largely centered near zero. Similarly, the density heatmap shows that forecast probabilities are tightly concentrated along the diagonal, confirming well-calibrated central estimates and demonstrating that the agreement is driven by the bulk of cases rather than isolated outliers.

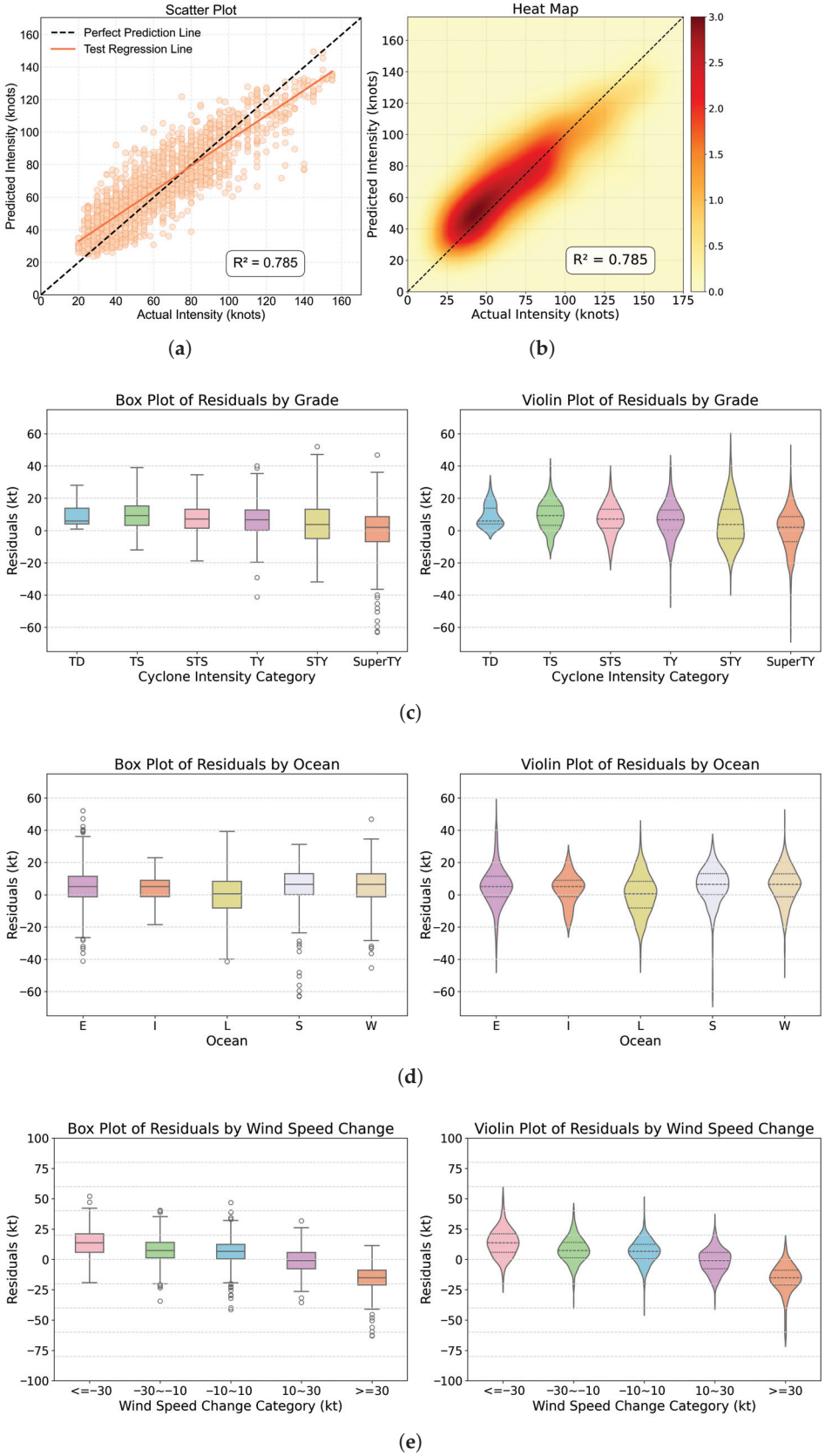


Figure 4. Comprehensive analysis of the prediction performance and uncertainty of the TCDM model. (**a**) Scatter plot of predicted versus observed TC intensity values. (**b**) Density heatmap of predictions. (**c**) Distribution of forecast residuals across different TC intensity grades. (**d**) Distribution of forecast residuals across ocean basins. (**e**) Distribution of forecast residuals across wind speed change.

Based on the tropical cyclone (TC) intensity classification system of the China Meteorological Administration (CMA), we performed a boxplot visualization of the absolute error distributions across different intensity categories, as shown in Figure 4c. The absolute error values were calculated as the algebraic differences between model outputs and observed values. Quantitative analysis indicates a pronounced negative correlation between TC intensity category and model accuracy: as the intensity increases from tropical depression (TD) and tropical storm (TS) to severe typhoon (STY) and super typhoon (SuperTY), the median absolute error exhibits a monotonic upward trend. The boxplots for TD and TS reveal compact distributions with a relatively low occurrence of outliers, suggesting that the model maintains higher predictive stability for low- to moderate-intensity cyclones, which account for 55% of the total samples. In sharp contrast, the error distribution for SuperTY demonstrates substantial divergence, highlighting greater uncertainty and error magnitude in the parameterization of dynamical processes for high-intensity TCs. It is noteworthy that discrete outlier points (gray dots outside the boxes) are detected across all intensity categories. Complementary violin plots further illustrate not only the range of values along the vertical axis, but also their distribution. Most intensity values cluster around 0 kt, while the distributions for TY, STY, and SuperTY span a wider range, indicating larger estimation variability within these categories. Conversely, the narrower distributions for TD, TS, and severe tropical storm (STS) suggest the smaller error dispersion and relatively stronger predictive capability of the model for these lower-intensity categories.

In addition, we analyzed the model's forecast errors for tropical cyclones occurring in different ocean basins, namely the Eastern Pacific (E), Indian Ocean (I), North Atlantic (L), Southern Hemisphere (S), and Western Pacific (W). As shown in Figure 4d, the absolute forecast errors of TC intensity in 2017 exhibit pronounced regional differences. Forecast errors over the Southern Hemisphere are generally lower, demonstrating better accuracy and stability, whereas both the North Atlantic and Western Pacific show higher medians and wider dispersions, with a larger number of extreme deviations. Several factors may contribute to these differences. First, the coverage and quality of observational data directly affect forecast performance. Although the North Atlantic benefits from a relatively mature observing system, its complex land–ocean boundaries and the frequent occurrence of rapid intensification events make it difficult for the model to accurately characterize TC intensity evolution. In contrast, the Southern Hemisphere, despite its more limited observational data, experiences lower cyclone activity and a relatively stable environmental background, which allows the model to better capture the intensity patterns and achieve lower forecast errors. Furthermore, geographical conditions, air–sea interaction characteristics, structural differences among cyclones, the accumulation of historical data, and the regional adaptability of the model also influence forecast accuracy. For example, the frequent occurrence of structurally complex and rapidly evolving super typhoons or hurricanes in the Western Pacific and North Atlantic increases the difficulty of intensity forecasting.

To further examine forecast performance under varying intensity change scenarios, samples were stratified according to the 24 h change in maximum sustained wind speed into five categories: strong weakening ($\Delta V < -30$ kt), moderate weakening ($-30 \leq \Delta V < -10$ kt), near-stationary ($-10 \leq \Delta V \leq 10$ kt), moderate strengthening ($10 < \Delta V \leq 30$ kt), and rapid intensification ($\Delta V > 30$ kt). As illustrated in Figure 4e, the absolute forecast errors vary systematically across these categories. The near-stationary group shows the lowest median error and the narrowest distribution, indicating that the model performs most reliably when cyclone intensity remains nearly unchanged. In contrast, both strong weakening and rapid intensification bins display substantially larger medians, broader dispersions, and more frequent extreme outliers, with the most pronounced errors occurring during rapid intensification events. These results reveal a clear dependence of error magnitude on the rate of intensity change. The elevated errors in the extreme-change categories are mainly attributable to two

factors. First, the physical processes governing rapid weakening or intensification—such as convective bursts, eyewall replacement cycles, ocean heat content anomalies, and abrupt air–sea feedbacks—are inherently nonlinear and difficult to resolve within the current model framework. Second, the scarcity of extreme-change cases in the training dataset reduces statistical robustness, leading to inflated variance and unstable estimation. Together, these factors explain the heteroscedastic error pattern, where forecast uncertainty grows with the absolute value of ΔV.

Collectively, these results indicate that while central predictions remain well calibrated, uncertainty bands should widen adaptively with intensity—a property naturally captured by TCDM's probabilistic formulation.

3.2. Comparison Model

Table 2 presents the performance of TCDM and the baseline models in cyclone intensity forecasting. To ensure a fair comparison across baseline models, all probabilistic approaches were evaluated under comparable computational resources. Specifically, the proposed TCDM used 120 ensemble members, a configuration identified as the performance saturation point according to the analysis presented in Figure 3. For comparison, the MC Dropout model performed 120 stochastic forward passes during inference, while the Deep Ensemble model consisted of 10 independently trained networks, each generating 12 samples, resulting in the same total of 120 ensemble members, and the GAN model was configured with 120 ensemble members. A comparison with baseline models indicates that TCDM achieves a balanced performance in both deterministic and probabilistic predictions. Our proposed model outperforms other machine learning approaches in the testing phase. To verify the significance of TCDM's performance improvements, we conducted the Diebold–Mariano statistical significance test. The results indicate that, compared with all other probabilistic benchmark models (MC Dropout, Deep Ensemble, and GAN), the improvements achieved by TCDM in terms of MAE, RMSE, and CRPS are statistically significant ($p < 0.05$).

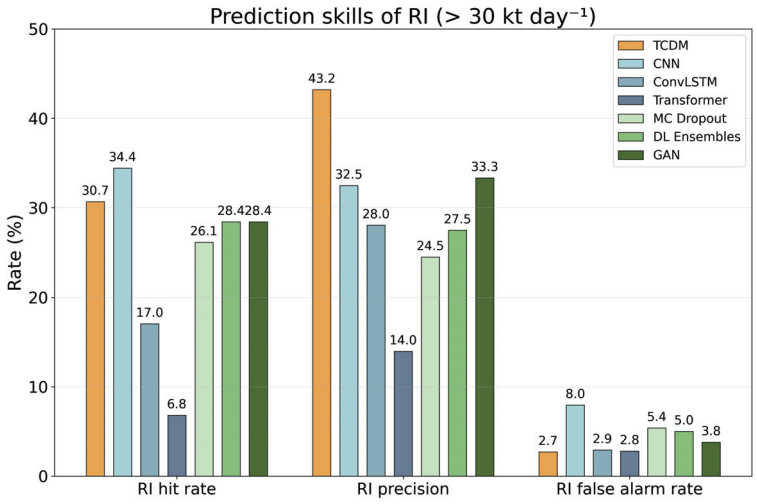
Table 2. Performance comparison between the TCDM model and baseline models.

	MAE	**RMSE**	**R^2**	**CRPS**	**PICP**
CNN	13.17	17.33	0.62	\	\
ConvLSTM	11.87	15.69	0.68	\	\
Transformer	11.71	14.93	0.71	\	\
MC Dropout	12.46	15.82	0.64	11.93	0.78
Deep Ensemble	12.11	15.78	0.66	9.34	0.81
GAN	11.21	14.24	0.73	7.98	0.86
TCDM	**10.04**	**12.73**	**0.78**	**7.17**	**0.93**

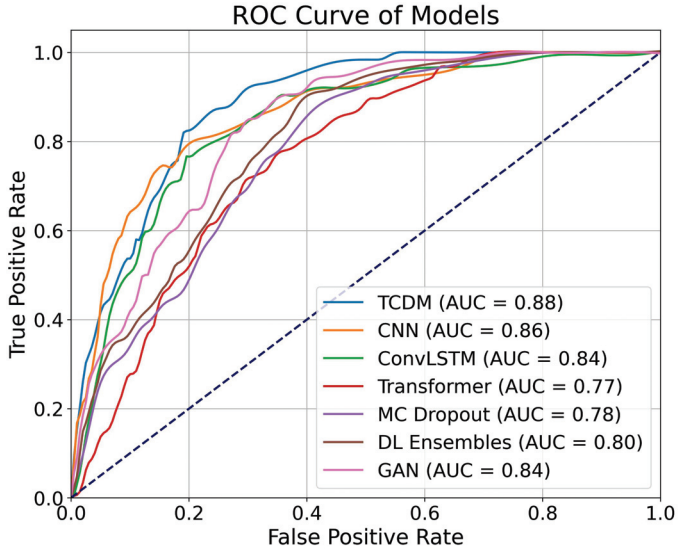
As shown in Table 2, the proposed TCDM model yields stable results for the CRPS metric in probabilistic forecasting. Notably, during independent testing, the 95% confidence interval of our model's PICP covers 93% of the test data, indicating superior interval prediction performance. The intervals nearly encompass the full range of TC intensity variations.

To evaluate the performance of the model in predicting rapid intensification (RI) phenomena, this study employed three commonly used classification metrics: precision (Precision = TP/(TP + FP)), hit rate (Hit Rate = TP/(TP + FN)), and false alarm rate (False Alarm Rate = FP/(FP + TN). Additionally, ROC curves for each model were plotted. The results shown in Figure 5a indicate that, at the set threshold, the TCDM model outperforms other models in both hit rate and precision, demonstrating its enhanced capability in RI event recognition. Further analysis reveals that the proposed method effectively suppresses false alarms, maintaining the false alarm rate below 5%, the lowest among all models. The

ROC curve and AUC values in Figure 5b further confirm the superiority of the TCDM model in rapid intensification prediction, with the curve closest to the top-left corner and the highest AUC, indicating its optimal overall performance in distinguishing rapid intensification events from non-rapid intensification events.



(**a**)



(**b**)

Figure 5. Comparative evaluation of rapid intensification (RI) prediction performance. (**a**) Model comparison based on precision, hit rate, and false alarm rate for RI events. (**b**) Receiver operating characteristic (ROC) curves and corresponding area under the curve (AUC) values.

3.3. Case Study

To further examine model behavior, Figure 6 presents probabilistic intensity forecasts for selected TCs. Overall, TCDM successfully captured the general evolution of cyclone intensity and provided uncertainty estimates that closely matched observed variability. The predictive distributions were generally compact and smooth, reflecting stable and reliable performance.

However, in some RI cases, while the model correctly signaled intensification trends, it tended to underestimate the upper bounds of extreme wind speeds. This bias likely stems from the relatively limited number of RI samples in the training set, which restricts the model's ability to fully learn the abrupt dynamics of such events. Despite this limitation,

the 50% and 80% confidence intervals encompassed most of the observed intensities, underscoring the model's capacity to express forecast uncertainty.

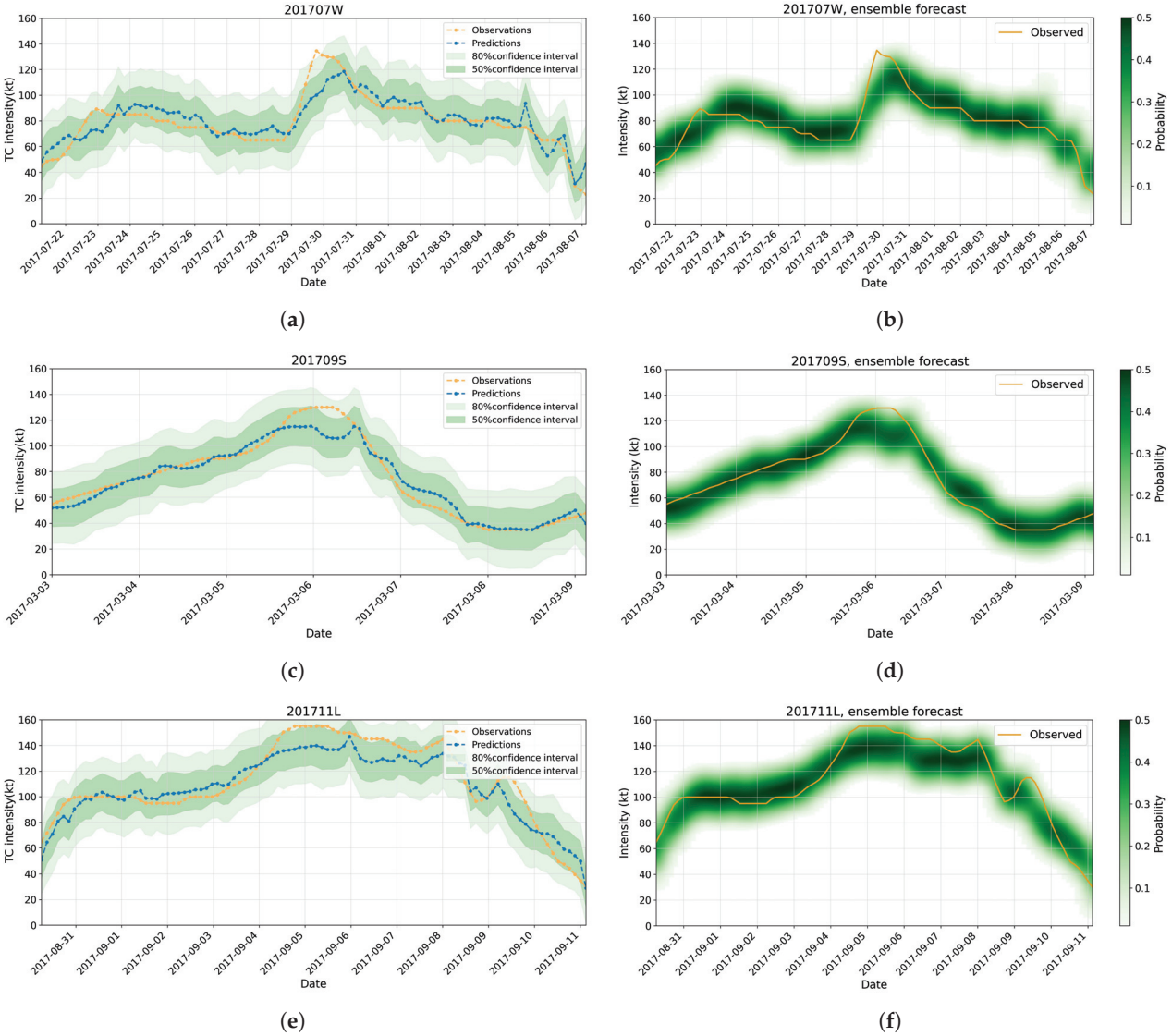


Figure 6. Probabilistic forecasts and confidence intervals for TCs 201707W, 201709S, and 201711L, generated by TCDM. (**a**,**c**,**e**): Time series of 24-h intensity forecasts for TC 201707W, TC 201709S, and TC 201711L, respectively. The plots show the observed intensity ("Observations"), the TCDM-predicted median intensity ("Predictions"), and the 50% (light green shaded area) and 80% (light blue shaded area) confidence intervals. (**b**,**d**,**f**): Corresponding full ensemble probabilistic forecasts for the three TCs. These plots visualize the probability density (heatmap) of the ensemble predictions at each time step, illustrating the evolution of the full predicted distribution and its uncertainty. The observed intensity ("Observed") is overlaid as a yellow line for comparison against the predicted probability distribution.

From a scientific perspective, these findings highlight both the promise and challenges of probabilistic TC forecasting. The strong performance in baseline scenarios underscores the value of diffusion-based probabilistic modeling, while the underestimation of extreme RI cases suggests a need for enhanced sample balancing or physics-guided constraints in future work.

4. Discussion

The TCDM model demonstrates significant advantages for the probabilistic forecasting of tropical cyclone intensity. Its core contribution lies in the successful introduction of diffusion generative models into the prediction of extreme meteorological events, enabling

explicit modeling of uncertainty. Compared with traditional deterministic forecasting methods, TCDM not only provides point predictions but also generates full posterior probability distributions, thereby offering a more comprehensive reflection of forecast risks and richer information for decision-making. The results indicate that TCDM achieves high hit rates and low false alarm rates in identifying RI events, suggesting strong capability in capturing nonlinear behaviors such as rapid intensity changes. This performance is primarily attributed to the diffusion model's strength in fitting complex conditional distributions and the effectiveness of the multimodal feature fusion mechanism.

Nevertheless, the TCDM model still has certain limitations. First, in cases of extreme rapid intensification, the model slightly underestimates the upper bound of intensity. This may be partially due to the insufficient proportion of RI samples in the training data and also reflects a limited sensitivity of the model to abrupt atmospheric thermodynamic processes. Second, the current model mainly relies on re-analysis data and remote sensing images as input, but it does not incorporate high-temporal-resolution and high-spatial-resolution in situ observations (e.g., surface heat fluxes or ocean mixed-layer temperatures), which may limit the representation of certain key physical processes. In addition, we conducted an exploratory experiment to evaluate the model's performance at longer forecast lead times. The results indicate a gradual degradation in both deterministic and probabilistic skill as lead time increases. Specifically, MAE and RMSE exhibit a near-linear growth with forecast horizon, while CRPS values increase more markedly beyond 96 h, suggesting a widening of predictive uncertainty. This performance decline is consistent with the general loss of predictive skill observed in both numerical and machine learning-based intensity forecasts at extended lead times, primarily due to the accumulation of environmental uncertainty and the increasingly stochastic nature of cyclone–ocean–atmosphere interactions. Nonetheless, even at 120 h, TCDM maintains lower error magnitudes and better-calibrated probabilistic intervals compared with baseline ensemble methods, highlighting its robustness and the benefits of diffusion-based uncertainty modeling for medium-range tropical cyclone forecasts. Furthermore, in terms of computational cost, the training of TCDM takes approximately 12 h on an A800 GPU. Generating 120 probabilistic forecast samples requires an average of 3.5 s, whereas the Deep Ensemble model takes only 0.4 s. This confirms the challenge of sampling efficiency inherent in diffusion models, which represents a key issue to be addressed in future work through model compression or accelerated sampling techniques.

These findings are consistent with recent studies that leverage deep generative models to improve meteorological forecasting [41], particularly demonstrating that probabilistic generative approaches offer more stable statistical performance than traditional ensemble methods when dealing with high-uncertainty events. However, compared with fully numerical model-based forecasting systems, purely data-driven methods still exhibit limitations in explaining certain abrupt weather processes due to the lack of explicit physical constraints [42]. Future work may focus on several directions: (1) incorporating learnable physical constraints (e.g., thermodynamic equations or energy balance conditions) into the model to enhance process plausibility; (2) integrating higher spatiotemporal resolution remote sensing retrieval parameters (e.g., sea surface temperature and liquid water path) to improve the perception of tropical cyclone microphysics and environmental features; (3) exploring model compression and accelerated inference techniques, such as knowledge distillation or sparse diffusion sampling, to enhance feasibility for operational applications.

The TCDM framework also demonstrates strong generalization and transfer potential and can be extended to probabilistic forecasting of additional meteorological variables, such as typhoon track ensemble prediction, heavy rainfall probability forecasting, and extreme weather diagnostics. In cross-disciplinary applications, including regional climate modeling and hydrometeorological coupled forecasting, the model shows promising applicability.

From an operational forecasting perspective, it can provide more scientific probabilistic support for meteorological disaster emergency management, particularly in early identification of rapid intensification events and quantification of associated uncertainties.

Finally, regarding the physical mechanisms learned by the model, we hypothesize that the cross-attention fusion module in TCDM can dynamically adjust its focus on satellite image features (e.g., convective organization structures) according to the environmental fields represented by SHIPS variables. For example, when the SHIPS data indicate the presence of strong vertical wind shear, the model may place greater emphasis on asymmetric structures in satellite cloud imagery. However, to bridge the gap between the "black-box" nature of the model and the meteorological understanding of underlying physical processes, future research should incorporate more in-depth interpretability analyses to explicitly identify the key multimodal features upon which the model relies during forecasting.

5. Conclusions

The TCDM model proposed in this study demonstrates strong accuracy and uncertainty quantification capabilities in probabilistic intensity forecasting of tropical cyclones. By incorporating diffusion modeling mechanisms and conditional information fusion, the model effectively captures complex relationships among meteorological variables, outperforming existing methods across multiple regions and intensity categories. In the rapid intensification detection task, TCDM achieves a well-balanced trade-off among accuracy, hit rate, and false alarm control, confirming its robustness in extreme weather scenarios. Furthermore, experiments reveal the distribution characteristics of model errors across different intensity levels and oceanic regions, further validating its generalization ability and practical application potential. Future work will focus on enhancing the model's capacity to learn from rare rapid intensification cases, improving predictions of moderate changes, and exploring the integration of additional physical priors with generative modeling strategies to advance the intelligence and reliability of disaster forecasting.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/rs17213600/s1, Supporting Information for "Probabilistic Forecasting Model for Tropical Cyclone Intensity Based on Diffusion Model".

Author Contributions: Conceptualization, J.L. and F.M.; methodology, P.Y.; software, P.Y.; validation, J.L., P.Y., and F.M.; formal analysis, J.L.; investigation, P.Y.; resources, F.M.; data curation, P.Y.; writing—original draft preparation, P.Y.; writing—review and editing, J.L. and F.M.; visualization, P.Y.; supervision, F.M.; project administration, J.L.; funding acquisition, J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the State Key Laboratory of Climate System Prediction and Risk Management (CPRM) initiative project (Grant No. CPRM-2025-NUIST-012), Climate System Prediction Research Center, the National Natural Science Foundation of China (Grant No. 62502219) and the Natural Science Foundation of Jiangsu Province (Grant No. BK20240700).

Data Availability Statement: The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. World Meteorological Organization. Rising Risks. 2022. Available online: https://public.wmo.int/en/resources/world-meteorological-day/world-meteorological-day-2022-early-warning-early-action/risking-risks (accessed on 4 September 2025).
2. Emanuel, K. Increasing destructiveness of tropical cyclones over the past 30 years. *Nature* **2005**, *436*, 686–688. [CrossRef]
3. Patricola, C.M.; Wehner, M.F. Anthropogenic increase in Atlantic hurricane intensities. *Science* **2018**, *359*, eaao6757.

4. Chen, B.-F.; Kuo, Y.-T.; Huang, T.-S. A deep learning ensemble approach for predicting tropical cyclone rapid intensification. *Atmos. Sci. Lett.* **2023**, *24*, e1151. [CrossRef]
5. Kaplan, J.; DeMaria, M. Large-scale characteristics of rapidly intensifying tropical cyclones in the North Atlantic basin. *Weather Forecast.* **2003**, *18*, 1033–1046. [CrossRef]
6. DeMaria, M.; Mainelli, M.; Shay, L.K.; Knaff, J.A.; Kaplan, J. The Hurricane Weather Research and Forecasting (HWRF) model. *Mon. Weather Rev.* **2005**, *133*, 1801–1816.
7. Cangialosi, J.P.; Blake, E.; DeMaria, M.; Penny, A.; Latto, A.; Rappaport, E.; Tallapragada, V. Recent progress in tropical cyclone intensity forecasting at the national hurricane center. *Weather Forecast.* **2020**, *35*, 1913–1922. [CrossRef]
8. DeMaria, M.; Knaff, J.A.; Sampson, C.R.; Musgrave, K.D.; Kaplan, J. Advances in tropical cyclone intensity forecasting. *Weather Forecast.* **2014**, *29*, 523–549.
9. Rogers, R.; Aberson, S.; Aksoy, A.; Annane, B.; Black, M.; Cione, J.; Dorst, N.; Dunion, J.; Gamache, J.; Goldenberg, S.; et al. NOAA's hurricane intensity forecasting experiment: A progress report. *Bull. Am. Meteorol. Soc.* **2013**, *94*, 859–882. [CrossRef]
10. DeMaria, M.; Kaplan, J. A statistical hurricane intensity prediction scheme (SHIPS) for the Atlantic basin. *Weather Forecast.* **1994**, *9*, 209–220. [CrossRef]
11. DeMaria, M.; Kaplan, J. An updated statistical hurricane intensity prediction scheme (SHIPS) for the Atlantic and eastern North Pacific basins. *Weather Forecast.* **1999**, *14*, 326–337. [CrossRef]
12. Roy, C.; Kovordanyi, R. Tropical cyclone track forecasting techniques—A review. *Atmos. Res.* **2012**, *104*, 40–69. [CrossRef]
13. Gneiting, T.; Raftery, A.E. Weather forecasting with ensemble methods. *Science* **2005**, *310*, 248–249. [CrossRef]
14. Hazelton, A.T. Improved Tropical Cyclone Rapid Intensification Forecasts in the Operational HAFS. *Weather Forecast.* **2023**, *38*, 1335–1354.
15. Zhang, J.A.; Marks, F.D. Progress and challenges in understanding and forecasting tropical cyclone rapid intensification. *Trop. Cyclone Res. Rev.* **2021**, *10*, 1–13.
16. Haupt, S.E.; Pasini, A. (Eds.). *Artificial Intelligence Methods in the Environmental Sciences*; Elsevier: Amsterdam, The Netherlands, 2020.
17. Reichstein, M.; Camps-Valls, G.; Stevens, B.; Jung, M.; Denzler, J.; Carvalhais, N.; Prabhat. Deep learning and process understanding for self-explaining earth system science. *Nature* **2019**, *566*, 195–204. [CrossRef]
18. Ham, Y.-G.; Kim, J.-H.; Luo, J.-J. Deep learning for multi-year ENSO forecasts. *Nature* **2019**, *573*, 568–572. [CrossRef]
19. Espeholt, L.; Agrawal, S.; Sønderby, C.; Kumar, M.; Heek, J.; Bromberg, C.; Gazen, C.; Hickey, J.; Bell, A.; Kalchbrenner, N. Skillful twelve hour precipitation forecasts using large context neural networks. *Nature* **2022**, *609*, 503–508.
20. Shi, X.; Chen, Z.; Wang, H.; Yeung, D.-Y.; Wong, W.-K.; Woo, W.-C. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In Proceedings of the NIPS'15: Proceedings of the 29th International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; pp. 802–810.
21. Chen, R.; Zhang, W.; Wang, X. Machine learning in tropical cyclone forecast modeling: A review. *Atmosphere* **2020**, *11*, 676. [CrossRef]
22. Xu, W.; Balaguru, K.; August, A.; Lalo, N.; Hodas, N.; DeMaria, M.; Judi, D. Deep learning experiments for tropical cyclone intensity forecasts. *Weather Forecast.* **2021**, *36*, 1453–1470. [CrossRef]
23. Wimmers, A.; Velden, C.; Cossuth, J.H. Using deep learning to estimate tropical cyclone intensity from satellite passive microwave imagery. *Mon. Weather Rev.* **2019**, *147*, 2261–2282. [CrossRef]
24. Cloud, K.; Rozoff, C.; Lewis, W.; Wimmers, A.; Velden, C.; Kossin, J. A deep learning model for rapid intensification of Atlantic hurricanes. *Weather Forecast.* **2019**, *34*, 1745–1761.
25. Su, H.; Wu, L.; Jiang, J.H.; Pai, C.-W.; Zhai, A.; Fovell, R.G. Applying satellite observations of tropical cyclone internal structures to rapid intensification forecast with machine learning. *Geophys. Res. Lett.* **2020**, *47*, e2020GL089102. [CrossRef]
26. Yang, Q.; Lee, C.Y.; Tippett, M.K. A long short-term memory model for global rapid intensification prediction. *Weather Forecast.* **2020**, *35*, 1203–1220. [CrossRef]
27. Meng, F.; Yang, K.; Yao, Y.; Wang, Z.; Song, T. Tropical Cyclone Intensity Probabilistic Forecasting System Based on Deep Learning. *Int. J. Intell. Syst.* **2023**, 3569538. [CrossRef]
28. Meng, F.; Yao, Y.; Wang, Z.; Peng, S.; Xu, D.; Song, T. Probabilistic forecasting of tropical cyclones intensity using machine learning model. *Environ. Res. Lett.* **2023**, *18*, 044042. [CrossRef]
29. Berner, J.; Achatz, U.; Batté, L.; De La Cámara, A.; Christensen, H.M.; Colangeli, M.; Colangeli, M.; Coleman, D.R.B.; Crommelin, D.; Dolaptchiev, S.I.; et al. Stochastic parameterization: Toward a new view of weather and climate models. *Bull. Am. Meteorol. Soc.* **2017**, *98*, 565–588. [CrossRef]
30. Gneiting, T.; Raftery, A.E. Probabilistic forecasts, calibration and sharpness. *J. R. Stat. Soc. Ser. Stat. Methodol.* **2007**, *69*, 243–268. [CrossRef]
31. Price, I.; Sanchez-Gonzalez, A.; Alet, F.; Andersson, T.R.; El-Kadi, A.; Masters, D.; Ewalds, T.; Stott, J.; Mohamed, S.; Battaglia, P.; et al. Probabilistic weather forecasting with machine learning. *Nature* **2025**, *637*, 84–90. [CrossRef] [PubMed]

32. Taillardat, M.; Mestre, O.; Zamo, M.; Naveau, P. Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics. *Mon. Weather Rev.* **2016**, *144*, 2375–2393. [CrossRef]
33. Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. In Proceedings of the NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 6–12 December 2020; pp. 6840–6851.
34. Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 2256–2265.
35. Song, Y.; Ermon, S. Improved techniques for training score-based generative models. In Proceedings of the NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 6–12 December 2020; pp. 12438–12448.
36. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 21–24 June 2022; pp. 10684–10695.
37. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In Proceedings of the Advances in Neural Information Processing Systems 27 (NIPS 2014), Montreal, QC, Canada, 8–13 December 2014; pp. 2672–2680.
38. Kingma, D.P.; Welling, M. Auto-encoding variational Bayes. In Proceedings of the 2nd International Conference on Learning Representations (ICLR), Banff, AB, Canada, 14–16 April 2014.
39. Chen, B.; Chen, B.-F.; Lin, H.-T. Rotation-blended CNNs on a new open dataset for tropical cyclone image-to-intensity regression. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), London, UK, 19–23 August 2018.
40. hariwal, P.; Nichol, A. Diffusion models beat GANs on image synthesis. In NIPS'21: Proceedings of the 35th International Conference on Neural Information Processing System, Virtual, 6–14 December 2021.
41. Ravuri, S.; Lenc, K.; Willson, M.; Kangin, D.; Lam, R.; Mirowski, P.; Fitzsimons, M.; Athanassiadou, M.; Kashem, S.; Madge, S.; et al. Skilful precipitation nowcasting using deep generative models of radar. *Nature* **2021**, *597*, 672–677. [CrossRef] [PubMed]
42. Beusch, L.; Rasp, S.; Dueben, P.D. Physics-constrained deep learning postprocessing of temperature and humidity. *Geophys. Res. Lett.* **2023**, *50*, e2023GL103582.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

big data and cognitive computing

MDPI

Article

Wildfire Prediction in British Columbia Using Machine Learning and Deep Learning Models: A Data-Driven Framework

Maryam Nasourinia and Kalpdrum Passi *

School of Engineering and Computer Science, Laurentian University, Sudbury, ON P3E 2C6, Canada; mnasourinia@laurentian.ca
* Correspondence: kpassi@laurentian.ca

Abstract

Wildfires pose a growing threat to ecosystems, infrastructure, and public safety, particularly in the province of British Columbia (BC), Canada. In recent years, the frequency, severity, and scale of wildfires in BC have increased significantly, largely due to climate change, human activity, and changing land use patterns. This study presents a comprehensive, data-driven approach to wildfire prediction, leveraging advanced machine learning (ML) and deep learning (DL) techniques. A high-resolution dataset was constructed by integrating five years of wildfire incident records from the Canadian Wildland Fire Information System (CWFIS) with ERA5 reanalysis climate data. The final dataset comprises more than 3.6 million spatiotemporal records and 148 environmental, meteorological, and geospatial features. Six feature selection techniques were evaluated, and five predictive models—Random Forest, XGBoost, LightGBM, CatBoost, and an RNN + LSTM—were trained and compared. The CatBoost model achieved the highest predictive performance with an accuracy of 93.4%, F1-score of 92.1%, and ROC-AUC of 0.94, while Random Forest achieved an accuracy of 92.6%. The study identifies key environmental variables, including surface temperature, humidity, wind speed, and soil moisture, as the most influential predictors of wildfire occurrence. These findings highlight the potential of data-driven AI frameworks to support early warning systems and enhance operational wildfire management in British Columbia.

Keywords: wildfire prediction; British Columbia; machine learning; deep learning; feature selection; Canadian wildland fire information system; climate reanalysis data; ERA5

1. Introduction

Wildfires have increasingly become one of the most critical global environmental challenges, threatening human communities, natural ecosystems, and biodiversity. Intensified by rising temperatures and drier atmospheric conditions associated with climate change, wildfire events are now occurring with greater frequency, severity, and spatial extent. According to Global Forest Watch, an average of 4.2 million hectares of forest are consumed annually by wildfires, incurring economic losses estimated at USD 20 billion globally. In addition to immediate destruction, wildfires significantly degrade air quality, disrupt ecological balance, and contribute to greenhouse gas emissions and global warming.

Traditional wildfire prediction systems, such as the Canadian Fire Weather Index (FWI) and statistical models, often rely on expert knowledge and linear assumptions. These methods struggle to capture the complex, nonlinear, and high-dimensional interactions between meteorological, topographical, and biological variables that drive wildfire behavior. The emergence of data-driven approaches—particularly machine learning (ML) and deep learning (DL) techniques—offers powerful tools capable of learning intricate patterns from multi-source environmental data. ML/DL algorithms are now increasingly applied to fire detection, forecasting, weather monitoring, and risk mapping.

Despite their potential, several limitations hinder the operational use of ML in wildfire prediction. These include the scarcity of high-resolution and reliable datasets, challenges in model interpretability, and difficulties in ensuring generalization across diverse geographic regions. In British Columbia (BC), where vast forests and diverse climates make wildfire management especially complex, these limitations are compounded by inconsistent data collection and limited integration across sources.

The present study is motivated by the urgent need for more accurate and interpretable wildfire prediction systems in BC, particularly in light of recent catastrophic fire seasons. In 2023, Canada experienced its most destructive wildfire year on record, with over 17 million hectares burned. In response, national initiatives have been launched to integrate AI and ML into fire risk assessment and resource allocation. This research contributes to those efforts by developing a comprehensive ML-based framework tailored to BC's specific environmental conditions.

The primary objective is to build and evaluate a high-performance wildfire prediction system that uses ML techniques and addresses critical challenges such as data imbalance, model transparency, and regional scalability. The framework involves integrating meteorological, geospatial, and fire incident data, applying advanced feature selection methods, and benchmarking the performance of various ML/DL models—including Random Forest, XGBoost, LightGBM, CatBoost, and RNN + LSTM.

Wildfires have become an increasingly pressing global concern, driven by rising temperatures, prolonged droughts, and heightened climate variability [1]. In Canada—particularly in British Columbia (BC)—wildfires have grown in frequency, scale, and intensity, causing severe environmental, economic, and societal impacts [2]. Accurate and timely prediction is therefore essential for informing early warning systems, optimizing emergency response, and supporting resource management strategies.

Despite substantial advances in wildfire modeling, existing studies often remain limited to small-scale regions or rely on coarse-resolution datasets, which restrict their operational applicability. This study distinguishes itself by developing a high-resolution, province-wide wildfire dataset for British Columbia, integrating over 3.6 million spatiotemporal records from multiple authoritative sources, including CWFIS, ERA5, FIRMS, and the British Columbia Wildfire Management Branch. In addition, the work presents a systematic comparison of six feature selection techniques and five machine learning and deep learning models, enabling a rigorous assessment of model performance and interpretability. By framing the results within the context of Canadian wildfire management, the research provides a scalable and data-driven decision-support framework designed to enhance both predictive capability and operational readiness for wildfire risk monitoring.

This study makes the following key contributions: (1) Development of a high-resolution wildfire dataset for British Columbia, constructed by integrating multi-source data, including the Canadian Wildland Fire Information System (CWFIS), ERA5 reanalysis data, and geospatial attributes. (2) Systematic comparison of six feature selection techniques (ReliefF, Relief, Mutual Information, Recursive Feature Elimination, Correlation-based, and Model-based Importance) across multiple ML and DL algorithms, providing

insight into their relative strengths for environmental prediction. (3) Comprehensive evaluation of five predictive models—Random Forest, XGBoost, LightGBM, CatBoost, and RNN + LSTM—using multiple train–test splits and 10-fold cross-validation to ensure generalizability. (4) Implementation of Random Undersampling (RUS) to address extreme class imbalance in wildfire data, improving model reliability and performance stability. Identification of key environmental drivers of wildfire occurrences, such as temperature, humidity, wind speed, and soil moisture, with strong alignment to established fire science findings. (5) Proposition of a scalable, interpretable, and region-specific ML/DL framework for wildfire prediction, designed to support data-driven decision-making warning systems in British Columbia.

2. Related Literature

Wildfires are a natural component of Canadian ecosystems and play a role in forest regeneration [1,2]. However, their frequency and intensity have increased in recent decades, driven by climate-induced shifts such as elevated temperatures, reduced snowpack, and persistent drought conditions [1,3,4]. National assessments indicate that Canada now experiences approximately 7000 wildfires annually, burning nearly 2.5 million hectares of land [5]. Human activity also plays a significant role, as nearly 55% of wildfires in Canada are caused by anthropogenic sources such as discarded cigarettes, unattended campfires, and industrial operations [6]. The 2016 Fort McMurray wildfire remains the costliest natural disaster in Canadian history, with over USD 9 billion in damages [7].

British Columbia accounts for nearly 40% of the total area burned annually in Canada [5,8]. The province's diverse topography and climate- ranging from humid coastal regions to dry interior plateaus- create highly variable and difficult-to-manage fire behavior [5,8]. Climatic trends have lengthened the fire season and intensified fire activity, with the 2021 season alone resulting in more than 8600 square kilometers burned [8]. Factors such as dry summers, strong winds, mountainous terrain, and climate-induced snowmelt contribute to BC's elevated wildfire risk [2,4,9]. In response, the BC Wildfire Service has implemented mitigation strategies that include real-time satellite surveillance, aerial patrols, FireSmart community initiatives, and partnerships with Indigenous communities [8,10,11].

Traditional wildfire prediction methods primarily rely on statistical and physical models. Statistical techniques such as logistic regression and Poisson models use historical data and meteorological variables—like temperature and humidity—to estimate fire probabilities [12]. These methods are simple and interpretable but generally fail to capture nonlinear relationships or complex interactions. Physical models, such as FARSITE, simulate fire spread using inputs such as terrain, fuel characteristics, and weather conditions [13]. Although effective for short-term fire behavior simulation, these models require detailed input data and are computationally intensive, limiting their use in real-time applications [14].

Remote sensing-based fire modeling approaches have been widely applied for detecting fire activity, mapping burn severity, and analyzing vegetation stress [15].

Remote sensing and GIS-based methods have expanded the capabilities of fire risk monitoring and assessment. Satellite systems like MODIS, Sentinel-2, and Landsat provide multispectral imagery for fire detection, vegetation stress analysis, and burn severity evaluation [16]. GIS platforms allow researchers to integrate spatial data layers—such as elevation, land cover, and human infrastructure—to assess fire susceptibility across regions [17]. However, these tools can be hampered by cloud cover, delayed data availability, and subjectivity in risk indicator selection.

Machine learning (ML) has become increasingly popular for wildfire prediction due to its ability to learn complex, high-dimensional relationships from diverse data sources. Supervised learning models—including Random Forest, Support Vector Machines, XGBoost,

and LightGBM—have demonstrated high predictive accuracy in identifying fire-prone areas [11,17,18]. Deep learning approaches such as Convolutional Neural Networks (CNNs) have been used for image-based fire detection [19], while Recurrent Neural Networks (RNNs) with Long Short-Term Memory (LSTM) units are applied to capture temporal patterns in weather data [20]. Hybrid models that combine spatial and temporal inputs or multiple classifiers further enhance robustness and generalizability [21].

Despite their strengths, ML models face several challenges. First, the lack of real-time, high-resolution data-especially in remote regions like BC- limits model effectiveness [16]. Second, generalization across geographic regions remains an issue, as many models are tuned to specific environments and fail under new conditions [21]. Third, datasets used for wildfire prediction often suffer from severe class imbalance, where non-fire events vastly outnumber actual fire occurrences [22]. Lastly, many ML models are considered "black boxes," offering little insight into their decision-making processes. This lack of transparency hinders their acceptance among policymakers and operational decision-makers [20].

Recent studies have attempted to overcome these challenges. Elsarrar et al. [23] used a rule-extraction neural network model for wildfire prediction in West Virginia, achieving high accuracy but with limited regional applicability. Hong et al. [21] implemented a hybrid method combining genetic algorithms with Random Forest and SVM models for fire susceptibility mapping in China, obtaining strong AUC scores but neglecting temporal dynamics. Omar [20] trained an LSTM-based model on a small Algerian dataset with high classification accuracy, though generalizability was not addressed. Bayat and Yildiz [24] compared ML models in Portugal and found SVM to perform best, while Sharma [25] identified Random Forest as the top model for Indian wildfire data. Li et al. [26] used ANN and SVM for forest fire prediction in China, finding ANN superior, though temporal components were not modeled.

Beyond conventional fire-science studies, interdisciplinary AI research has increasingly linked machine learning to broader environmental-resilience and sustainability frameworks. Olawade et al. [27] emphasized that AI-driven environmental monitoring enables proactive ecosystem management and disaster risk reduction. Camps-Valls et al. [28] demonstrated how deep learning models can detect and predict extreme climate events—including wildfires—with greater spatial precision. Forouheshfar et al. [29] discussed the integration of AI and ML into adaptive-governance structures that enhance systemic resilience to climate change, while Wu et al. [30] illustrated ML's capability to optimize sustainability in complex systems such as green supply chains. Collectively, these interdisciplinary insights establish that wildfire prediction frameworks grounded in machine learning not only advance technical performance but also contribute to data-driven environmental governance and climate adaptation.

By synthesizing insights from climate science, AI research, and sustainable policy design, the present study situates wildfire prediction as part of a wider global effort toward resilient, adaptive, and intelligent environmental systems, offering both methodological rigor and operational value for Canadian wildfire management.

Building on these prior works, the present study introduces a localized, high-resolution dataset for British Columbia, integrates both spatial and temporal predictors, and employs a hybrid feature-selection strategy alongside Random Undersampling (RUS) to address class imbalance. The study also leverages deep learning architectures—specifically RNNs with LSTM units—to capture temporal dependencies and applies a comprehensive multi-metric evaluation framework to ensure robustness and interpretability. These contributions collectively advance the development of transparent, data-driven, and operationally scalable wildfire forecasting systems.

3. Materials and Methods

3.1. Dataset and Data Preprocessing

This study utilizes a comprehensive dataset developed specifically for wildfire prediction in British Columbia (BC), Canada. Due to the province's ecological diversity and vulnerability to climate change, wildfires have become increasingly frequent and destructive. Constructing a high-quality dataset is thus essential for building reliable predictive models that reflect the spatial, temporal, and environmental conditions affecting fire behavior in this region [1,2,6].

The dataset combines multiple sources. Historical wildfire records were collected from the Canadian Wildland Fire Information System (CWFIS) [31] and the British Columbia Wildfire Management Branch [10]. These datasets include precise geolocation, ignition date, fire size, and cause. Meteorological data were sourced from Environment Canada [32] and Copernicus ERA5 reanalysis datasets, which offer daily values for temperature, humidity, wind speed, surface pressure, and precipitation. Global fire detection records from NASA's Fire Information for Resource Management System (FIRMS) [7] and land cover layers from Global Forest Watch [5] were also incorporated to improve spatial completeness and validation.

The data integration workflow is summarized in Figure 1. The process began with preprocessing of meteorological variables, which were originally stored in NetCDF format. These files were converted into CSV to facilitate data processing. Fire incident records were filtered by location and date, converted from shapefiles to structured tables, and aligned with climate data using spatial and temporal joins. A provincial boundary shapefile from Statistics Canada was applied to ensure that all observations fell within BC. Water body masks were used to remove points over lakes, rivers, and oceans, ensuring that fire labels would not be falsely assigned in non-burnable areas.

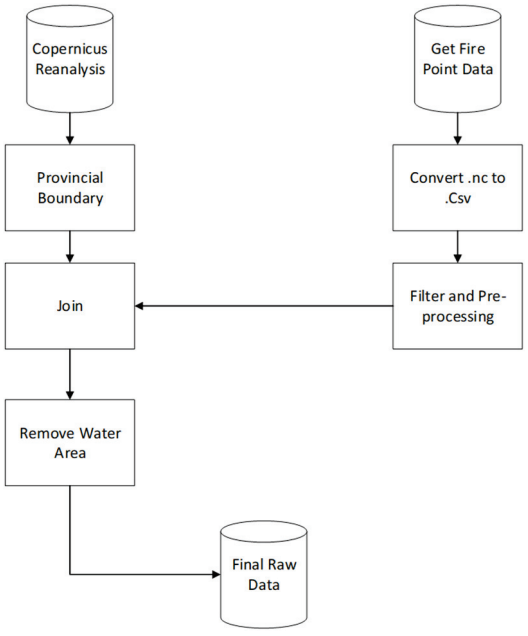


Figure 1. Dataset Creation Framework.

To assign binary classification labels, each spatial grid cell was labeled as 'fire' (1) or 'non-fire' (0) based on its proximity to a documented fire event on the same date. The Haversine distance formula was used to determine the proximity between ERA5 grid points and fire ignition records, ensuring accurate spatial linkage. The final dataset contains 3,631,455 rows and 148 features, covering a wide range of meteorological, geographic, and

temporal variables. A summary of the overall dataset composition and its characteristics is presented in Table 1, compiled from multiple data sources including CWFIS [31], FIRMS [7], Environment Canada [32], BC Wildfire Service [10], and Global Forest Watch [5].

Table 1. Dataset Overview (compiled from [5,7,10,31,32]).

Component	Description
Total Rows	3,631,445
Total Features	148
Target Classes	2 (Binary classification: fire/no fire)

In addition to standard weather indicators, the dataset includes derived features such as vegetation indices, soil moisture, and evapotranspiration rates, which are important predictors of fire risk. Geospatial attributes such as latitude, longitude, and elevation add spatial context, while calendar variables such as day-of-year and seasonality help model cyclical fire behavior. This rich feature space supports the training of both conventional and deep learning models.

To enhance predictive performance and reduce overfitting, six feature selection techniques were applied independently. These included Relief, ReliefF, correlation-based filtering, recursive feature elimination (RFE), mutual information (MI), and model-based selection using Random Forest and XGBoost [18]. Each method was configured using empirically optimized parameters and applied to the complete dataset. The selected feature subsets were then used to train separate models, and the results were compared using accuracy, recall, F1-score, and ROC-AUC. The configuration settings used for each feature selection method are presented in Table 2. The parameter settings summarized in Table 2 were empirically optimized to achieve a balance between predictive accuracy and computational efficiency. For the Relief and ReliefF algorithms, the neighbor values (k = 10–20) were determined experimentally to ensure stable feature ranking while maintaining manageable runtime. The correlation threshold of ± 0.8 was adopted following previous wildfire-related studies [18,22] to minimize multicollinearity without removing informative predictors. The number of selected features (30–60) was tuned to capture the most relevant meteorological and geospatial variables while preventing model overfitting. These parameter choices were validated across multiple random splits to ensure consistency.

To provide further insight into the behavior of each method, a qualitative comparison of their strengths, selection strategies, and motivation is included in Table 3. This comparison helped guide the selection of the most effective combinations of feature selection and modeling techniques for wildfire classification.

A major challenge in wildfire prediction is the severe class imbalance between fire and non-fire observations. In this dataset, the overwhelming majority of records were labeled as non-fire. To mitigate bias toward the dominant class, Random Undersampling (RUS) was applied. RUS randomly discards a portion of the majority class to equalize class frequencies, resulting in a balanced dataset. While alternative strategies such as SMOTE and SMOTE-ENN were considered, memory and runtime limitations in the Google Colab environment necessitated the use of a more efficient method [22].

Despite its simplicity, RUS significantly improved model sensitivity and reduced training time. However, it carries the trade-off of discarding potentially informative samples, which may affect generalization. Given these constraints, RUS was adopted as a practical balancing strategy for all models developed in this study.

Table 2. Parameter settings for feature-selection algorithms.

Method	Parameter	Value/Description
Mutual Information	Number of Features to Select	30–50 (based on information gain ranking)
	Discretization Strategy	Equal-width binning (default in mutual_info_classif)
	Random State	0
Relief	Number of Neighbors (k)	10
	Number of Features Selected	40 (top-ranked by average relevance scores)
ReliefF	Number of Neighbors (k)	20
	Distance Metric	Manhattan Distance (L1 norm)
	Number of Features Selected	55
Recursive Feature Elimination (RFE)	Estimator	Logistic Regression (or Random Forest)
	Number of Features Selected	30
	Step Size	1
Model-Based Selection	Base Estimator	Random Forest/XGBoost
	Importance Threshold	Mean importance > 0.005
	Number of Features Selected	~25
Correlation-Based	Pearson Correlation Threshold	± 0.8
	Removal Strategy	Drop one of each highly correlated feature pair
	Final Feature Count	~60

Table 3. Comparison of feature selection methods.

Method	Type	Strengths	Why It Was Selected
ReliefF	Instance-based	Handles noisy data, considers interactions	Extension of Relief; handles multiclass/noisy data better; considers more neighbors
Relief	Instance-based	Fast, simple	Fast filtering method that ranks features based on local relevance and handles noise well
Correlation-based	Statistical	Easy to implement, removes redundancy	Simple, efficient, and helps eliminate redundant features by analyzing linear correlations.
Model-based	Model-driven	Works with complex models, considers interactions	Uses feature importance from Random Forest; aligned with model behavior.
RFE	Wrapper	Finds optimal subset, improves model performance	Identifies and removes highly redundant or collinear features.
Mutual Information	Information-theoretic	Captures non-linear dependencies	Its ability to detect non-linear relationships without relying on a specific model

3.2. Methodology

This study proposes a comprehensive methodology for wildfire occurrence prediction in British Columbia, using a combination of advanced machine learning (ML) and deep learning (DL) models. The growing frequency and severity of wildfires, exacerbated by climate change and human interventions [1,2], necessitate the development of predictive systems capable of modeling complex environmental interactions. Building upon the

dataset described in the previous section, this framework incorporates stages of feature selection, class balancing, model training, and multi-metric evaluation.

Feature selection was conducted using a hybrid approach that combined filter-based methods such as ReliefF and correlation analysis, wrapper-based techniques like recursive feature elimination, and embedded approaches based on model-driven importance ranking [16,22]. Class imbalance, a known challenge in wildfire datasets, was addressed using Random Undersampling (RUS), which reduced the dominance of non-fire instances while preserving the integrity of test sets [11].

A variety of classifiers were evaluated, including ensemble models -Random Forest, XGBoost, LightGBM, and CatBoost-due to their proven robustness on structured environmental datasets [18,22]. In addition, a temporal deep learning architecture based on Recurrent Neural Networks (RNN) with Long Short-Term Memory (LSTM) units was incorporated to capture sequential weather patterns and improve modeling of time-dependent fire dynamics [3].

The complete modeling pipeline is illustrated in Figure 2. It begins with preprocessing and balancing the training data using RUS, followed by normalization via Min-Max scaling applied solely to the training data and later used to transform the test data, thereby preventing data leakage. Three different train–test split ratios were tested—90:10, 80:20, and 70:30—as well as 10-fold cross-validation, ensuring model robustness and generalizability. These evaluation strategies allowed a comparative assessment under varying training scenarios.

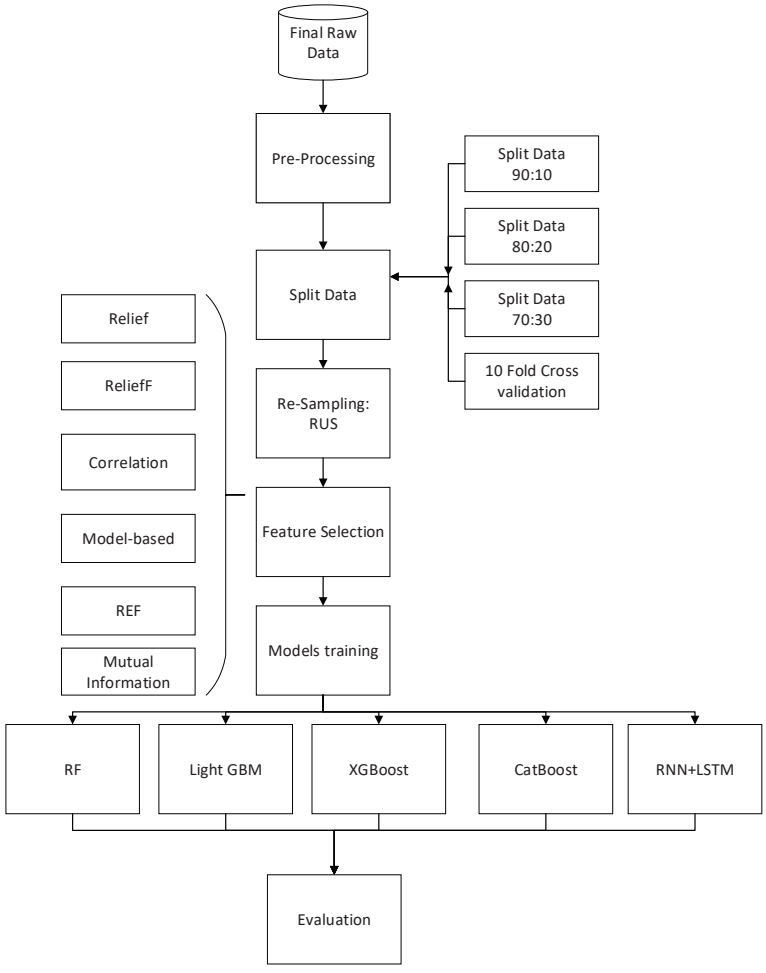


Figure 2. Schematic representation of the wildfire prediction modeling framework.

Hyperparameters for each model were selected based on empirical tuning, prior literature, and framework-specific best practices. These configurations are summarized in Table 4. For Random Forest, XGBoost, LightGBM, and CatBoost, parameters such as number of trees, learning rates, and maximum depth were adjusted for optimal performance. For RNN + LSTM, parameters included the number of LSTM units, hidden layers, optimizer type, and training epochs.

Table 4. Final hyperparameter configurations for all models.

Model	Hyperparameter	Value
Random Forest	Number of Trees (n_estimators)	100
	Maximum Tree Depth (max_depth)	10
	Minimum Samples per Split (min_samples_split)	2
XGBoost	Learning Rate (learning_rate)	0.1
	Number of Estimators (n_estimators)	200
	Maximum Depth (max_depth)	6
	Subsample Ratio (subsample)	0.8
	Column Sample per Tree (colsample_bytree)	0.8
CatBoost	Iterations (iterations)	500
	Learning Rate (learning_rate)	0.05
	Tree Depth (depth)	6
	Loss Function (loss_function)	Logloss
LightGBM	Number of Leaves (num_leaves)	200
	Learning Rate (learning_rate)	0.05
	Number of Estimators (n_estimators)	100
RNN + LSTM	LSTM Units per Layer	64
	Number of Hidden Layers	2
	Optimizer	Adam
	Learning Rate	0.001
	Epochs	20
	Batch Size	128

All hyperparameters in Table 4 were tuned via grid search with five-fold cross-validation on the balanced training set, optimizing primarily F1 and AUC to reflect minority-class detectability. We constrained the search to configurations that remain computationally feasible for near–real-time use. For Random Forest, n_estimators = 100 provided stable estimates while keeping inference latency low; max_depth = 10 limited variance and prevented overfitting on highly correlated meteorological features; min_samples_split = 2 preserved tree diversity after balancing.

For XGBoost, a conservative learning_rate = 0.1 with n_estimators = 200 achieved strong AUC without excessive training time. max_depth = 6 balanced non-linear expressiveness with generalization; subsample = 0.8 and colsample_bytree = 0.8 injected stochasticity to reduce overfitting and improve robustness across spatial–temporal splits. Early stopping on validation AUC was applied during tuning. For CatBoost, iterations = 500 and learning_rate = 0.05 yielded the best bias–variance trade-off; depth = 6 controlled model complexity on tabular data with mixed feature interactions. The Logloss objective aligns with binary classification while permitting calibrated probabilities used downstream in threshold analysis. For LightGBM, num_leaves = 200 allowed finer partitioning than max_depth while exploiting histogram-based splits; learning_rate = 0.05 and n_estimators = 100 stabilized training and reduced overfitting under early stopping. For RNN + LSTM, two hidden layers with 64 units each were sufficient to capture intra-seasonal dynamics (multi-day lags) without over-parameterization. We used Adam with learning_rate = 0.001 for stable

optimization across heterogeneous scales, batch size = 128 for efficient GPU utilization, and 20 epochs based on validation loss saturation; longer training offered negligible gains but increased variance.

These settings were validated across random, temporal (train: earlier seasons; test: later seasons), and cross-regional splits, and consistently delivered strong F1/AUC with manageable runtime and memory footprints suitable for operational wildfire-risk workflows.

Among the machine learning models, Random Forest provided a reliable baseline due to its capacity for handling high-dimensional data and resistance to overfitting [17,18]. XGBoost, using gradient boosting with regularization and second-order optimization, offered strong performance by sequentially minimizing prediction error through additive tree construction [21,23]. CatBoost, with its native handling of categorical data and ordered boosting mechanism, provided efficiency and resilience to overfitting even in mostly numerical datasets [20]. LightGBM, known for its leaf-wise tree growth and histogram-based feature splitting, exhibited rapid training times and high predictive accuracy, particularly suitable for high-dimensional, imbalanced data [24].

The RNN + LSTM architecture enabled temporal modeling by ingesting sequences of meteorological data, learning dependencies across time, and producing predictions based on historical patterns. This architecture was especially effective for detecting wildfire onset driven by evolving climatic conditions [25,26].

Model performance was evaluated using five complementary metrics: accuracy, recall (sensitivity), specificity, weighted F1-score, and ROC-AUC. Accuracy provided an overall measure of correctness but could be misleading in imbalanced datasets. Recall focused on the model's ability to correctly detect fire events, whereas specificity assessed false-positive reduction. The F1-score balanced precision and recall, and ROC-AUC captured class separability across decision thresholds. These evaluation metrics allowed for a holistic comparison and supported the selection of the most suitable model for real-world wildfire forecasting.

In conclusion, the methodology integrated robust data preparation, balanced model training, and multi-metric evaluation to ensure generalizable and interpretable predictions. This framework forms the foundation for experimental results and discussion presented in the following chapter.

4. Results

This section presents a detailed evaluation of the machine learning and deep learning models developed for wildfire prediction in British Columbia. The analysis spans multiple experimental setups, including various train–test splits (90:10, 80:20, 70:30), six feature selection techniques, and 10-fold cross-validation. Performance metrics—accuracy, recall, precision, F1 score, and ROC-AUC—were used to assess and compare model behavior under different conditions.

The evaluation began with performance analysis under the 90:10 train–test split. This configuration, providing the largest training data volume, enabled high predictive performance across most models. As detailed in Table 5 and visualized in Figures 3–7, XGBoost combined with Relief achieved the highest F1 score (0.997), while RNN + LSTM paired with correlation-based features yielded the best recall (0.992). CatBoost and Random Forest also demonstrated strong results, particularly with model-based feature selection, reflecting a good balance between precision and sensitivity. LightGBM showed excellent precision in several scenarios but occasionally lagged in recall, suggesting limited effectiveness in identifying rare fire events.

Table 5. Performance comparison of all models under the 90:10 train–test split; best results per metric (Accuracy, Precision, Recall, F1, AUC) are in bold.

Feature Selection	Model	Accuracy	Recall	Precision	F1 Score	AUC
Mutual_information						
	CatBoost	0.9749	0.904834	0.690462	**0.868525**	0.842998
	LightGBM	0.9562	0.601817	0.758961	0.774668	0.689962
	RNN + LSTM	0.8627	**0.907982**	0.842346	0.621196	0.968291
	RandomForest	**0.987**	0.697308	0.718856	0.799906	**0.977199**
	XGBoost	0.9719	0.824132	**0.880177**	0.705922	0.771469
RFE						
	CatBoost	0.9877	0.795813	0.662883	0.760288	0.861263
	LightGBM	0.9716	0.677038	**0.94058**	0.734333	**0.970451**
	RNN + LSTM	0.9626	**0.836977**	0.768763	0.644877	0.617252
	RandomForest	**0.9909**	0.645635	0.85642	**0.761164**	0.7942
	XGBoost	0.9866	0.620665	0.627675	0.701849	0.798134
Relief						
	CatBoost	0.9757	0.833417	**0.986704**	0.712684	**0.922498**
	LightGBM	0.9552	0.928791	0.875948	0.673784	0.902367
	RNN + LSTM	0.944	0.603453	0.781352	0.9165	0.800418
	RandomForest	**0.9845**	0.748153	0.778282	0.868606	0.848217
	XGBoost	0.9731	**0.987458**	0.671557	**0.997086**	0.830719
ReliefF						
	CatBoost	0.9758	0.701262	0.694178	**0.86214**	**0.820574**
	LightGBM	0.9547	**0.951275**	0.899849	0.704672	0.675592
	RNN + LSTM	0.9467	0.880528	**0.908273**	0.742597	0.67446
	RandomForest	**0.9827**	0.804954	0.681128	0.807457	0.705805
	XGBoost	0.9728	0.88598	0.705635	0.758421	0.769526
Correlation-based						
	CatBoost	0.6232	0.785464	0.780027	0.782738	0.609248
	LightGBM	0.6049	0.969095	0.910732	0.770365	0.668147
	RNN + LSTM	0.5917	**0.992045**	**0.983173**	**0.965933**	0.64258
	RandomForest	**0.6863**	0.691334	0.610652	0.949527	**0.936957**
	XGBoost	0.6209	0.914792	0.608008	0.80327	0.766213
Model-based						
	CatBoost	0.9888	**0.962328**	0.777761	0.633384	**0.897283**
	LightGBM	0.9699	0.743395	**0.996742**	0.788375	0.655742
	RNN + LSTM	0.9635	0.610133	0.982966	0.766718	0.706197
	RandomForest	**0.9914**	0.931887	0.847256	**0.806455**	0.683052
	XGBoost	0.9871	0.846523	0.838083	0.638342	0.770376

Table 5 presents the complete numerical performance metrics (Accuracy, Precision, Recall, F1-Score, and ROC-AUC) for all evaluated models under the 90:10 train–test split. These values allow direct quantitative comparison of algorithms.

In contrast, Figures 3–7 provide complementary visual analyses: Figure 3 compares accuracy distributions, Figure 4 displays ROC curves and AUC behavior, Figure 5 illustrates confusion matrices highlighting false negatives and false positives, and Figures 6 and 7 summarize feature importance and cross-model comparison. Together, these visualizations enhance interpretability beyond the numeric metrics summarized in Table 5.

Under the 80:20 split (Table 6), performance slightly declined due to reduced training data. Nonetheless, Random Forest and CatBoost maintained their lead, with Random Forest achieving 0.9897 accuracy (with RFE) and CatBoost reaching 0.9866 (model-based). XGBoost with ReliefF recorded an impressive F1 score of 0.958, confirming its stability across splits.

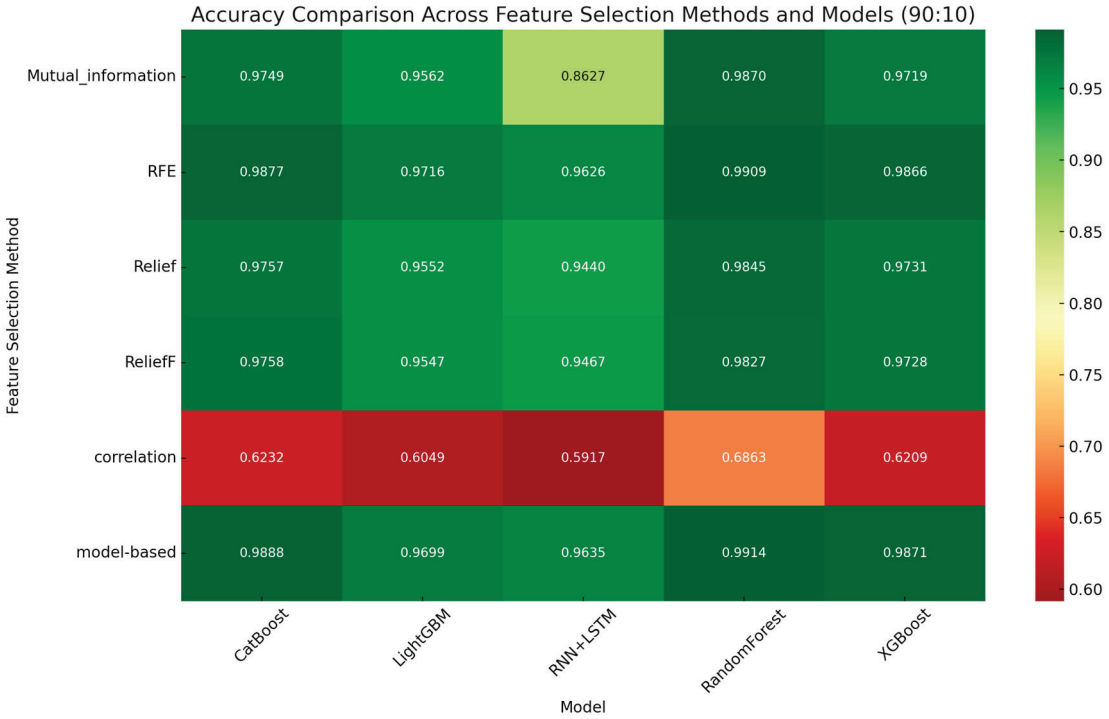


Figure 3. Accuracy comparison across models under the 90:10 split.

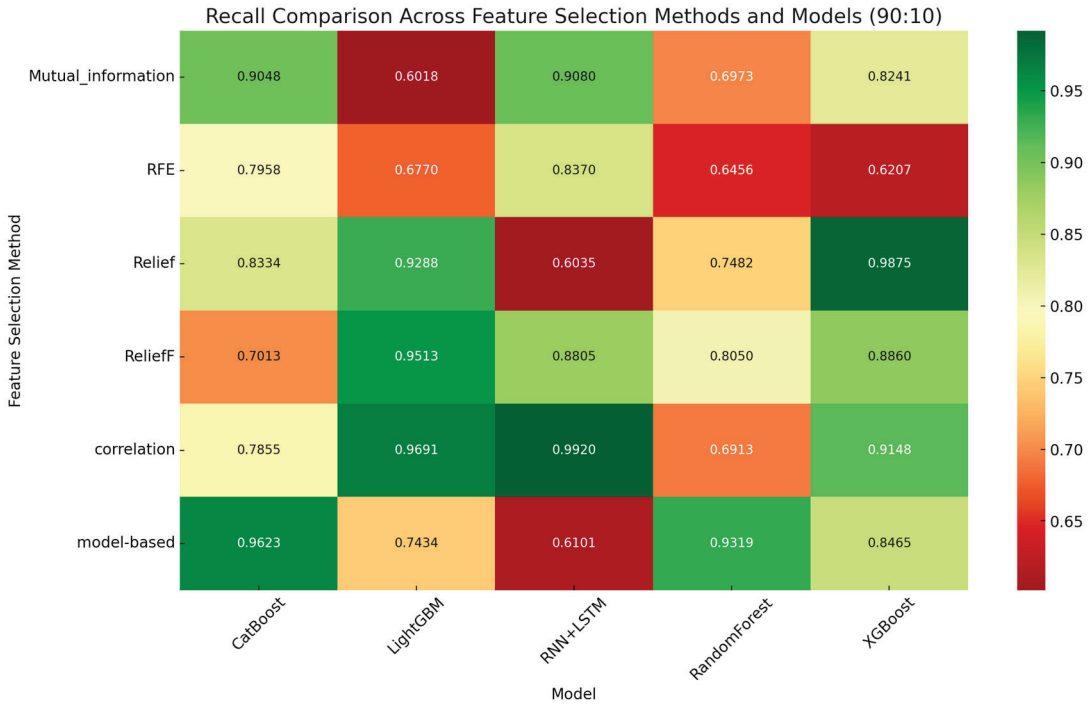


Figure 4. Recall Comparison 90:10.

The AUC performance of all models under the 80:20 train–test split is illustrated in Figure 8.

The most challenging configuration, the 70:30 split, provided further insight into model resilience (Table 7). Random Forest again delivered strong performance with 0.9894 accuracy and 0.984 F1 score (model-based). XGBoost excelled in recall (0.992) with Relief, while CatBoost remained competitive with multiple feature selectors. Deep learning, particularly RNN + LSTM, showed performance improvements under ReliefF and correlation-based features but remained sensitive to feature subset quality.

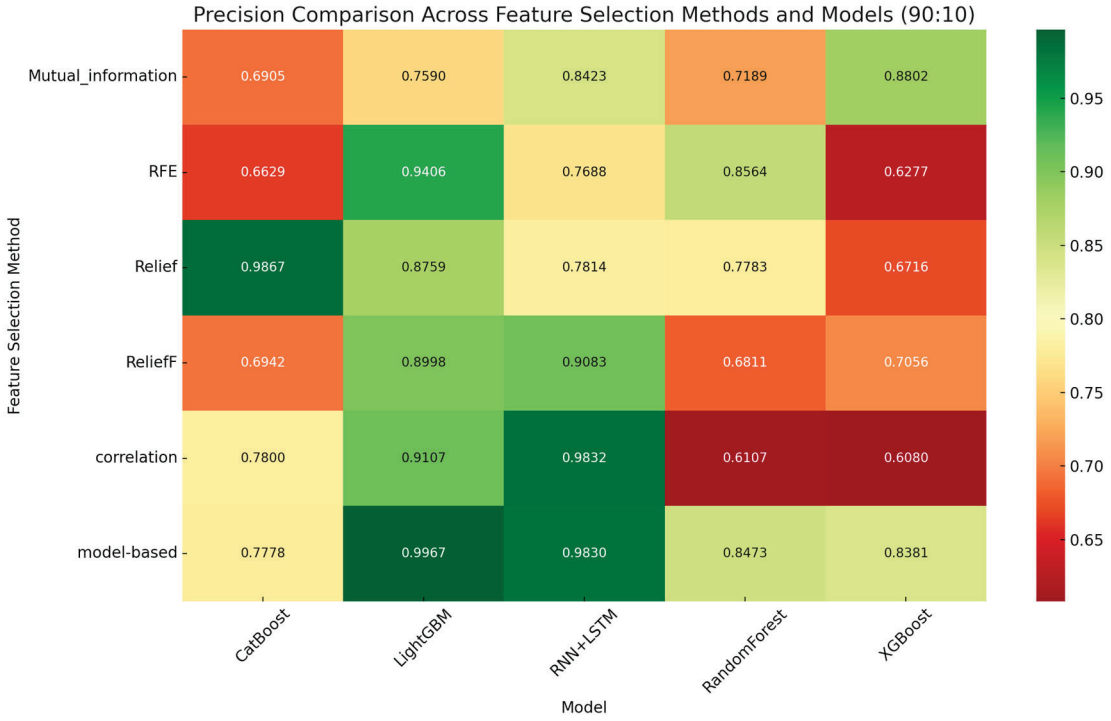


Figure 5. Precision Comparison 90:10.

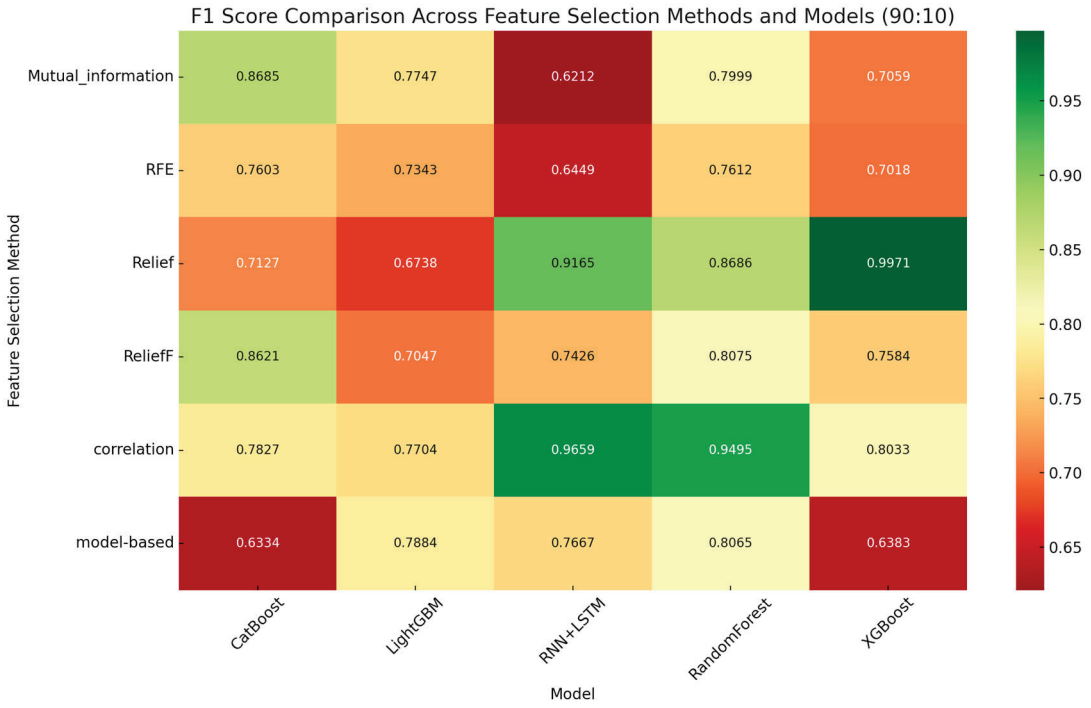


Figure 6. F1-Score Comparison 90:10.

To ensure robust conclusions, 10-fold cross-validation was applied (Table 8). Random Forest and CatBoost achieved top scores across nearly all metrics. Notably, Random Forest with Mutual Information reached 0.9818 recall and 0.9729 AUC, while CatBoost with Relief yielded a superior F1 score of 0.991. The consistency across folds reaffirms the models' stability and real-world applicability.

Figures 9–13 summarize the results of the 10-fold cross-validation for all evaluated models. As shown in Figure 9, the accuracy comparison demonstrates that ensemble-based models, particularly CatBoost and Random Forest, consistently achieved the highest

accuracy values across folds. Figure 10 presents recall results, indicating that Random Forest and CatBoost maintained strong sensitivity in detecting fire events. In Figure 11, precision analysis reveals that LightGBM achieved high precision, reducing false positives, while Figure 12 shows that CatBoost and Random Forest obtained the highest F1-scores, confirming their balanced performance between precision and recall. Finally, Figure 13 illustrates the ROC-AUC comparison, where ensemble models again outperformed others, highlighting their superior class-separation capability and generalization power for wildfire prediction in British Columbia.

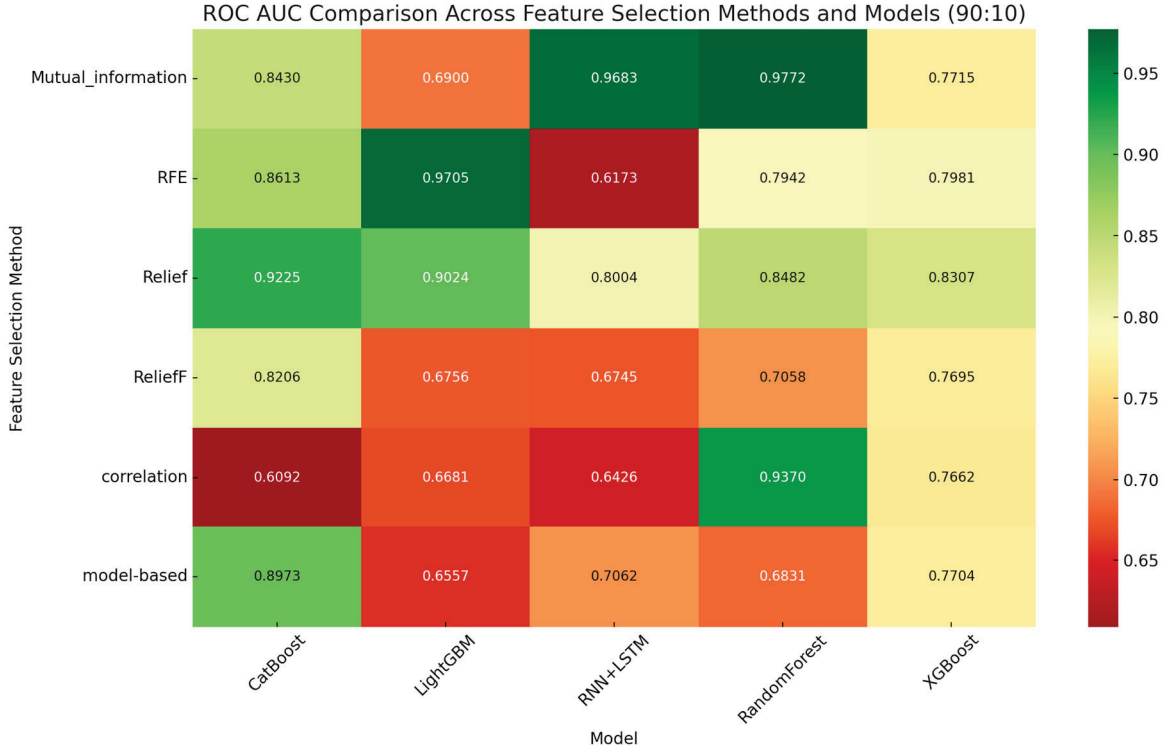


Figure 7. ROC-AUC Comparison 90:10.

Table 6. Results-80:20 Split.

Feature Selection Method	Model	Accuracy	Recall	Precision	F1 Score	AUC
Mutual Information						
	CatBoost	0.9739	0.798923	0.619009	0.805752	**0.964411**
	LightGBM	0.9529	0.975961	0.863294	0.742278	0.699431
	RNN + LSTM	0.8497	0.819235	0.80698	0.759353	0.956144
	RandomForest	0.9861	0.711722	0.911739	0.76679	0.825793
	XGBoost	0.9716	0.870813	0.947071	0.884573	0.862082
RFE						
	CatBoost	0.9857	0.649877	0.704808	0.8979	0.689285
	LightGBM	0.9702	0.805591	0.799191	0.643816	0.64842
	RNN + LSTM	0.9640	0.793378	0.901545	0.941477	0.632972
	RandomForest	0.9897	0.695229	0.811982	0.978512	0.924495
	XGBoost	0.9857	0.737249	0.85378	0.828005	0.639923

Table 6. *Cont.*

Feature Selection Method	Model	Accuracy	Recall	Precision	F1 Score	AUC
Relief						
	CatBoost	0.9730	0.885159	0.95307	0.770535	0.979375
	LightGBM	0.9525	0.766013	0.820477	0.662833	0.889025
	RNN + LSTM	0.9375	0.7	0.861508	0.65715	0.803496
	RandomForest	0.9810	0.640606	0.874684	0.609023	0.664299
	XGBoost	0.9713	0.654724	0.760414	0.827304	0.964871
ReliefF						
	CatBoost	0.9731	0.831297	0.841859	0.934922	0.846046
	LightGBM	0.9537	0.723748	0.968789	0.886612	0.603536
	RNN + LSTM	0.9459	0.797748	0.795881	0.719749	0.856571
	RandomForest	0.9787	0.76864	0.687718	0.767006	0.764384
	XGBoost	0.9708	0.831816	0.892801	0.958404	0.722814
Correlation-based						
	CatBoost	0.6138	0.915354	0.685669	0.807834	0.66691
	LightGBM	0.6004	0.637558	0.858328	0.736278	0.957402
	RNN + LSTM	0.5910	0.881865	0.883029	0.774056	0.964639
	RandomForest	0.6772	0.874622	0.765682	0.601329	0.804171
	XGBoost	0.6096	0.685832	0.756527	0.853353	0.903777
Model-based						
	CatBoost	0.9866	0.963091	0.632288	0.786762	0.889891
	LightGBM	0.9694	0.933881	0.708122	0.806424	0.726937
	RNN + LSTM	0.9624	0.882569	0.775223	0.66675	0.616749
	RandomForest	0.9897	0.814609	0.755561	0.932214	0.89216
	XGBoost	0.9849	0.914273	0.647892	0.850143	0.604667

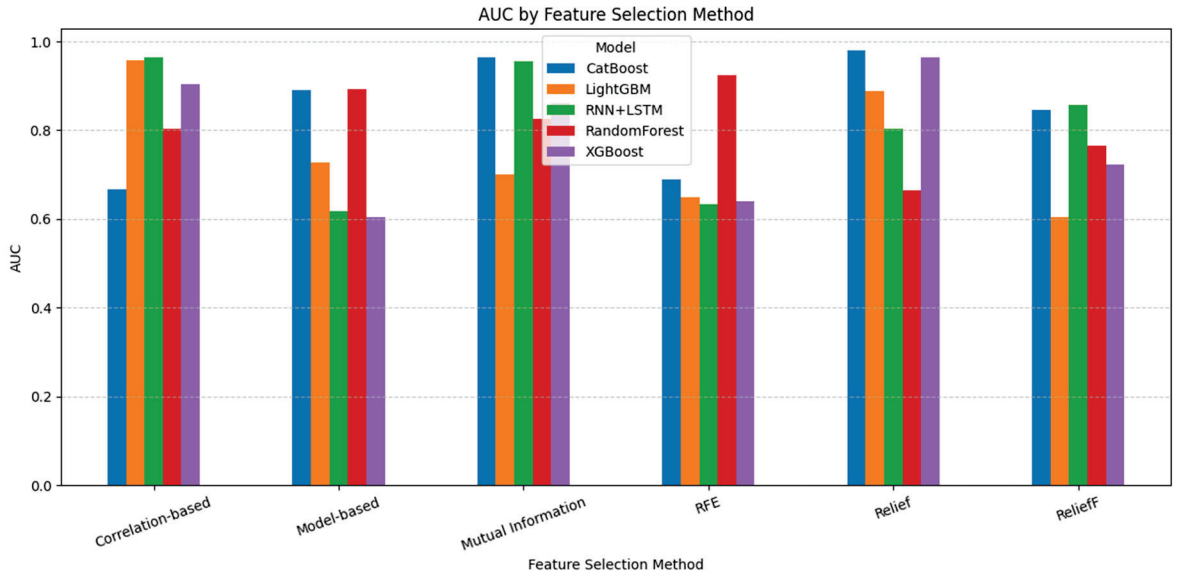


Figure 8. AUC performance 80:20.

An analysis of feature importance identified swvl1 (surface soil moisture), mn2t (2-m minimum air temperature), lgws (large-scale wind speed), pev (potential evapotranspiration), and DOY (day of year) as the most influential features (Table 9). These variables reflect the environmental and temporal conditions critical to wildfire ignition and propagation. Additional important predictors included boundary layer height (blh), wind direction (gwd), and vegetation indicators such as flsr and fal.

Table 7. Results 70:30 Split.

Feature Selection Method	Model	Accuracy	Recall	Precision	F1 Score	AUC
Mutual Information						
	CatBoost	0.9731	0.601151	0.912792	0.691087	0.85596
	LightGBM	0.9558	0.704806	0.923149	0.926469	0.951541
	RNN + LSTM	0.8501	0.764621	0.702396	0.71165	0.82491
	RandomForest	0.9851	0.760119	0.899627	0.740562	0.889904
	XGBoost	0.9722	0.88979	0.796421	0.754351	0.948969
RFE						
	CatBoost	0.9864	0.855432	0.625921	0.697662	0.765965
	LightGBM	0.9713	0.656193	0.851219	0.970671	0.852759
	RNN + LSTM	0.9626	0.919289	0.709009	0.635736	0.942664
	RandomForest	0.9894	0.666482	0.993943	0.809336	0.714497
	XGBoost	0.9861	0.835971	0.807732	0.772012	0.826582
Relief						
	CatBoost	0.9717	0.954848	0.836371	0.648531	0.84763
	LightGBM	0.9509	0.984228	0.674332	0.88741	0.802683
	RNN + LSTM	0.9384	0.664084	0.874892	0.979518	0.870642
	RandomForest	0.9802	0.816757	0.684596	0.622357	0.859792
	XGBoost	0.9707	0.992105	0.675924	0.682691	0.740771
ReliefF						
	CatBoost	0.9725	0.924294	0.860407	0.668035	0.663806
	LightGBM	0.9507	0.990708	0.614532	0.8353	0.662955
	RNN + LSTM	0.9385	0.710726	0.634973	0.824327	0.996142
	RandomForest	0.9785	0.647736	0.678948	0.817843	0.649228
	XGBoost	0.9706	0.773456	0.698027	0.792471	0.921085
Correlation-based						
	CatBoost	0.6126	0.949141	0.693303	0.785164	0.710548
	LightGBM	0.5995	0.932876	0.752983	0.730375	0.659205
	RNN + LSTM	0.5889	0.887946	0.907728	0.914135	0.633926
	RandomForest	0.6688	0.732814	0.937343	0.814342	0.681625
	XGBoost	0.6089	0.742217	0.652592	0.797131	0.868122
Model-based						
	CatBoost	0.9861	0.92615	0.858515	0.789538	0.726312
	LightGBM	0.9706	0.767083	0.991497	0.807526	0.923869
	RNN + LSTM	0.9622	0.649916	0.855903	0.810854	0.797215
	RandomForest	0.9887	0.721168	0.977737	0.984446	0.684643
	XGBoost	0.9856	0.645382	0.905245	0.89354	0.951307

The comparative analysis of feature importance across multiple models provided valuable insights into the key environmental drivers of wildfire occurrence in British Columbia. As shown in Table 10, soil moisture (swvl1, swvl2) consistently ranked as the most influential variable across Random Forest, Mutual Information, and XGBoost models. Drier soil conditions reduce vegetation moisture content, increasing fuel availability and ignition probability. Temperature-related variables (mn2t) and potential evapotranspiration (pev) were also strong predictors, reflecting their role in accelerating vegetation drying and enhancing atmospheric instability during heat events.

Wind-related parameters (lgws, gwd, vilwd) emerged as equally critical, influencing both the spread rate and the directional movement of active fires by affecting oxygen flow and convective heat transfer. Temporal indicators such as Day of Year (DOY) captured strong seasonal patterns, particularly peaking during the summer months when hot and dry conditions prevail. Atmospheric and boundary-layer factors such as blh (Boundary

Layer Height) further contributed to model accuracy by representing vertical heat exchange and atmospheric turbulence associated with fire spread dynamics.

Table 8. Ten-Fold Cross-Validation Results.

Feature Selection Method	Model	Accuracy	Recall	Precision	F1 Score	ROC AUC
Correlation	RandomForest	0.983265	0.97942	0.982101	0.953074	0.968299
Correlation	XGBoost	0.95423	0.964737	0.95511	0.949881	0.963224
Correlation	LightGBM	0.952397	0.943204	0.946739	0.936505	0.961408
Correlation	CatBoost	0.966788	0.965268	0.956541	0.977141	0.979601
Correlation	RNN + LSTM	0.956173	0.960005	0.957684	0.962904	0.961174
Model-Based	RandomForest	0.958425	0.972705	0.987196	0.965316	0.973554
Model-Based	XGBoost	0.949914	0.937312	0.947752	0.942976	0.962613
Model-Based	LightGBM	0.964294	0.943742	0.942688	0.966252	0.933951
Model-Based	CatBoost	0.951081	0.988495	0.97293	0.9792	0.95585
Model-Based	RNN + LSTM	0.938502	0.946088	0.967277	0.950001	0.951065
Relief	RandomForest	0.952675	0.958656	0.964963	0.977401	0.982945
Relief	XGBoost	0.958123	0.953603	0.957261	0.959296	0.936463
Relief	LightGBM	0.953043	0.935569	0.966182	0.937608	0.939073
Relief	CatBoost	0.966951	0.95236	0.971278	0.991197	0.954475
Relief	RNN + LSTM	0.960016	0.959566	0.929303	0.961941	0.964307
ReliefF	RandomForest	0.965352	0.97867	0.958083	0.972737	0.967559
ReliefF	XGBoost	0.946665	0.969082	0.943658	0.957563	0.943444
ReliefF	LightGBM	0.945846	0.958076	0.965361	0.959538	0.959496
ReliefF	CatBoost	0.965762	0.950695	0.971574	0.954171	0.979059
ReliefF	RNN + LSTM	0.950435	0.942902	0.938249	0.950994	0.928936
RFE	RandomForest	0.958625	0.976047	0.966811	0.981644	0.975979
RFE	XGBoost	0.94771	0.953999	0.94818	0.942667	0.940498
RFE	LightGBM	0.935693	0.946494	0.936696	0.961345	0.970275
RFE	CatBoost	0.981377	0.953284	0.964033	0.989793	0.954406
RFE	RNN + LSTM	0.931446	0.955465	0.937474	0.956263	0.937656
Mutual Information	RandomForest	0.974076	0.981811	0.977506	0.970682	0.972957
Mutual Information	XGBoost	0.95654	0.941941	0.962239	0.961731	0.966264
Mutual Information	LightGBM	0.964519	0.954746	0.941361	0.961991	0.960191
Mutual Information	CatBoost	0.963608	0.965012	0.968076	0.98322	0.965882
Mutual Information	RNN + LSTM	0.93827	0.964822	0.926848	0.945622	0.938091

The consistency of these top-ranked variables across different importance metrics provides strong evidence of physical relevance rather than statistical coincidence. Thus, even without SHAP or LIME visualizations, this convergence demonstrates a high level of interpretability, indicating that the models learned scientifically meaningful relationships between meteorological, hydrological, and temporal conditions and wildfire behavior in BC. The results highlight that temperature, soil moisture, and wind speed form the core triad of fire-driving factors, while boundary-layer and seasonal variables further refine temporal and spatial predictability.

Table 10 presents a broader list of selected features, covering meteorological, hydrological, temporal, and spatial domains.

Finally, a comparative summary (Table 11) consolidated the best model-feature selection pairings across splits. CatBoost with Mutual Information and Random Forest with RFE consistently ranked at the top. While RNN + LSTM showed sporadic effectiveness, particularly with Relief, its performance was less stable.

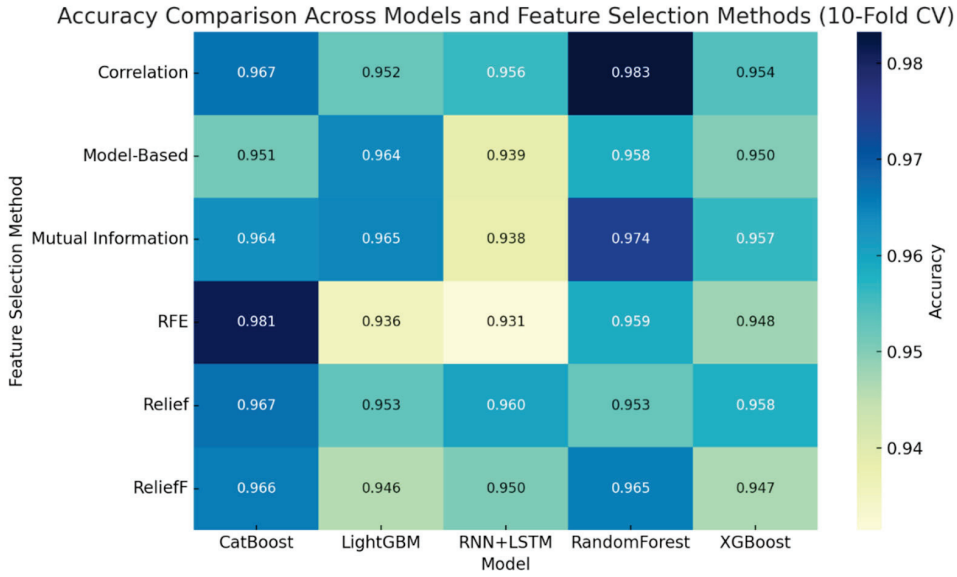


Figure 9. Accuracy in 10-Fold CV.

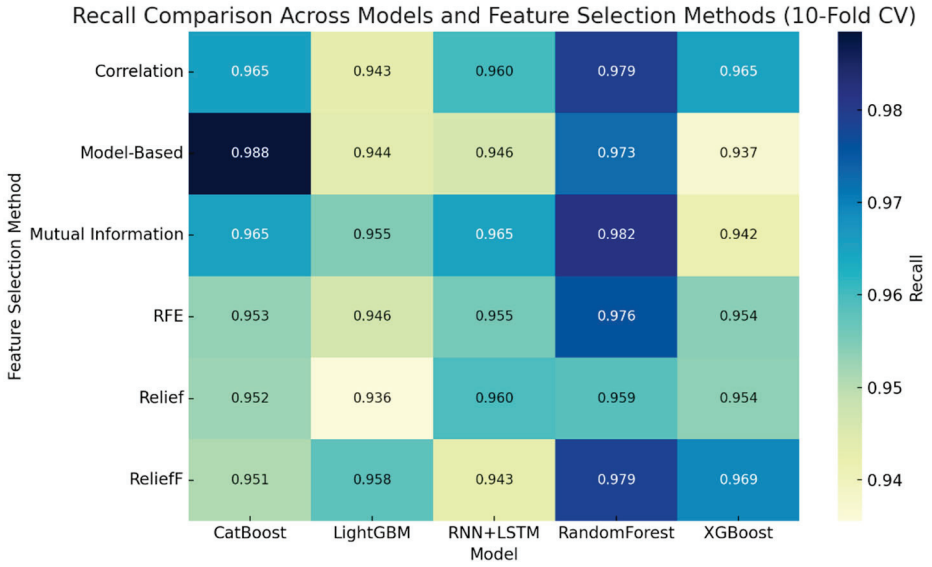


Figure 10. Recall in 10-Fold CV.

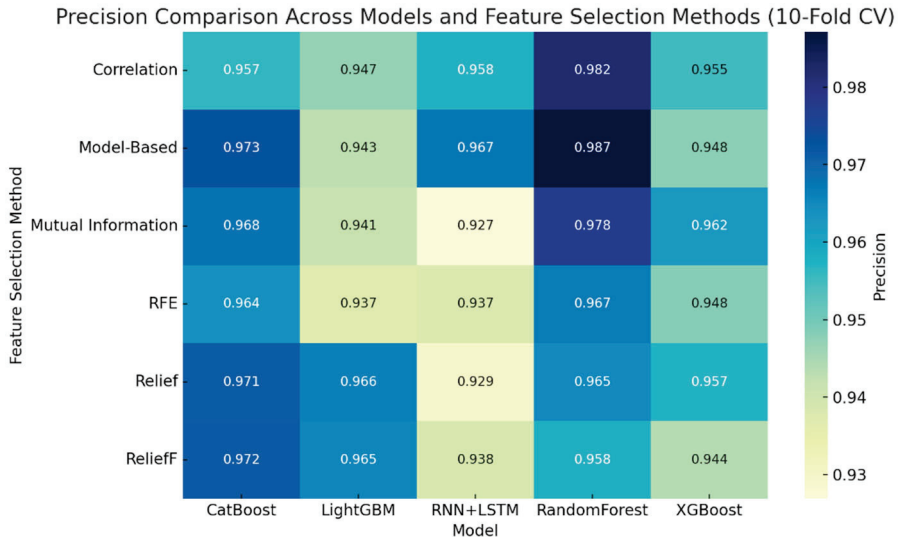


Figure 11. Precision 10-Fold CV.

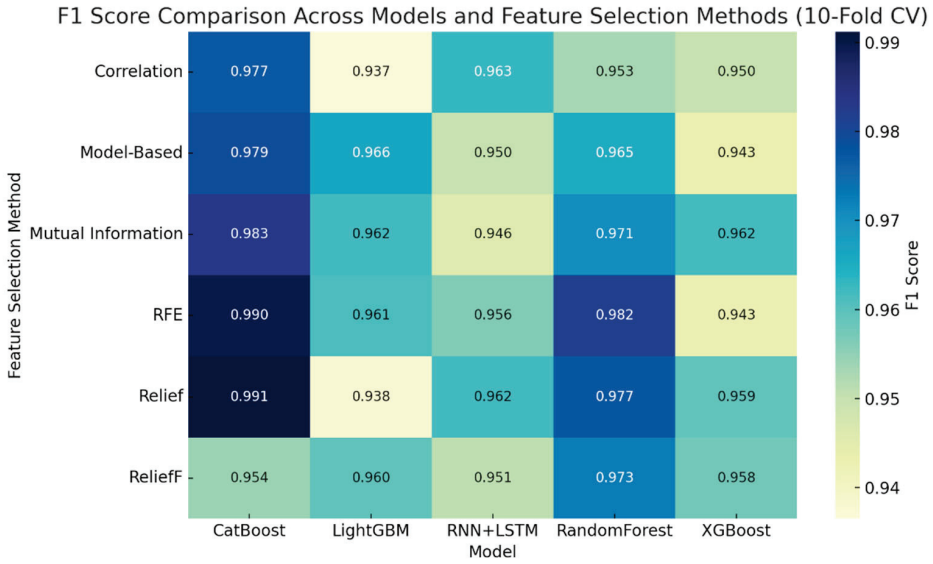


Figure 12. F1-Score results in 10-Fold CV.

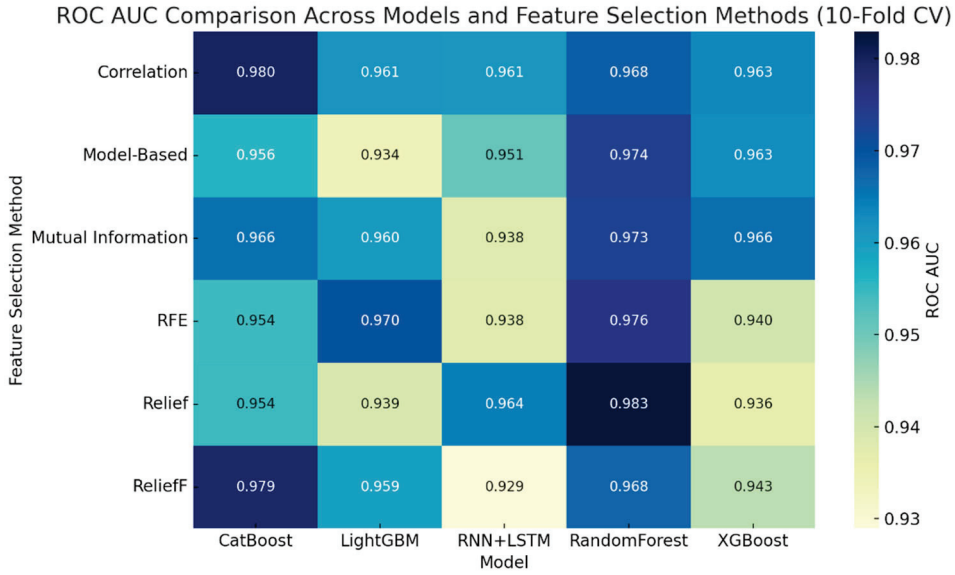


Figure 13. ROC-AUC Results 10-Fold CV.

Table 9. Top-10 Features.

Rank	Feature Name	Description	Importance Sources
1	swvl1	Surface soil moisture (layer 1)	RF, MI, XGBoost, ReliefF
2	mn2t	Minimum 2 m air temperature	RF, MI, XGBoost, Correlation
3	lgws	Large-scale wind speed	MI, ReliefF, XGBoost
4	pev	Potential evapotranspiration	RF, MI, XGBoost
5	DOY	Day of year (seasonality indicator)	Relief, MI, XGBoost
6	gwd	Wind direction	RF, XGBoost
7	blh	Boundary layer height	MI, XGBoost
8	mgws	Medium-scale wind speed	Relief, XGBoost
9	vilwd	Divergence of wind	RF, XGBoost
10	swvl2	Soil moisture (layer 2)	MI, RF

Table 10. Selected Features.

Feature Code	Full Name	Description
istl2	Instantaneous Soil Temperature Level 2	Soil temperature at a specific depth in the ground.
index_instant	Instantaneous Fire Index	Fire index based on real-time weather data.
blh	Boundary Layer Height	Height of the atmospheric boundary layer, affecting fire behavior.
vilwd	Vertical Integrated Liquid Water Density	Measures moisture content in the atmosphere.
fal	Fractional Land Cover	Percentage of land covered by vegetation, crucial for fuel assessment.
vilwe	Vertical Integrated Liquid Water Equivalent	Measures water content in clouds, affecting precipitation.
viiwe	Vertically Integrated Ice Water Equivalent	Indicator of ice particles in the atmosphere, relevant for cloud and precipitation dynamics.
es	Evaporation Stress Index	Represents water stress on vegetation, affecting fire spread.
gwd	Geostrophic Wind Direction	Wind flow pattern at high altitudes, influencing fire movement.
lgws	Large-Scale Wind Speed	Measures strong winds that impact fire intensity.
nsss	North-South Surface Stress	Measures frictional force of wind along the north-south axis.
DOY	Day of Year	Temporal feature used to track seasonal fire patterns.
ttr	Total Top Radiation	Solar radiation received at the top of the atmosphere, influencing fire ignition.
ttrc	Total Cloud Cover Radiation	Measures cloud influence on radiation, affecting temperature and humidity.
deg0l	Zero-Degree Level Height	Altitude where temperature is 0 °C, affecting precipitation type (rain vs. snow).
mn2t	Minimum 2 m Air Temperature	Lowest daily temperature at 2 m above ground.
pev	Potential Evapotranspiration	Amount of water evaporated and transpired, affecting soil dryness.
flsr	Fraction of Land Surface Reflectance	Indicator of vegetation health and surface dryness.
asn	Accumulated Snow	Total snow accumulation, influencing moisture levels.
vilwn	Vertical Integrated Liquid Water Northward	Measures northward movement of moisture in the atmosphere.
viiwd	Vertical Integrated Ice Water Downward	Measures downward motion of ice water, affecting precipitation.
ie	Instantaneous Evaporation	Measures real-time evaporation, affecting soil moisture.
viiwn	Vertical Integrated Ice Water Northward	Tracks northward movement of ice particles in the atmosphere.
index	General Fire Index	A calculated index representing overall fire risk.
index_max	Maximum Fire Index	The highest fire index value observed in a given period.
stl2	Soil Temperature Level 2	Temperature of soil at a deeper layer than istl2.
bld	Boundary Layer Depth	Measures the depth of the boundary layer, affecting heat and moisture exchange.
mgws	Mean Gust Wind Speed	Measures peak wind gusts, influencing rapid fire spread.
swvl1	Soil Water Volume Level 1	Measures water content in the topsoil layer, affecting vegetation dryness.
chnk	Convective Heating Near Surface	Tracks heat transfer near the ground, influencing local fire conditions.

Table 11. Comparative analysis.

Split	Best for Accuracy	Best for Recall	Best for Precision	Best ROC AUC
90:10	0.9914 (Model-based + RF)	0.9920 (Correlation-based + RNN + LSTM)	0.9967 (Model-based + LightGBM)	0.9644 (Mutual Info + CatBoost)
80:20	0.9897 (RFE + RF)	0.9760 (Mutual Info + LightGBM)	0.9967 (Model-based + LightGBM)	0.9794 (Relief + CatBoost)
70:30	0.9894 (RFE + RF)	0.9921 (Relief + XGBoost)	0.9939 (RFE + Random Forest)	0.9961 (ReliefF + RNN + LSTM)

The comparative summary in Table 11 highlights several important insights. Ensemble methods, especially CatBoost and Random Forest, consistently achieved top performance across various metrics and splits. Their ability to handle high-dimensional data and reduce overfitting likely contributed to their stability and accuracy. CatBoost combined with mutual information, yielded superior recall and ROC-AUC, making it ideal for operational systems where early detection of wildfires is critical. Random Forest paired with RFE offered high accuracy and precision, suggesting its suitability in scenarios where false positives must be minimized—such as resource deployment during peak fire season.

Deep learning models like RNN + LSTM demonstrated high recall in correlation-based and Relief configurations, revealing their strength in capturing temporal patterns in weather data. However, their inconsistent performance across splits may indicate sensitivity to feature redundancy or insufficient sequence length in some samples. Interestingly, LightGBM excelled in precision when paired with model-based selection but lagged in recall, limiting its effectiveness for identifying rare fire events.

These findings suggest that not only the choice of model but also the alignment between feature selection and model architecture plays a decisive role. Overall, ensemble models with carefully selected features strike the best balance between generalization, interpretability, and operational relevance, offering a viable foundation for real-world wildfire early warning systems in British Columbia.

Among all evaluated algorithms, the ensemble-based models—particularly CatBoost and XGBoost—consistently achieved superior predictive performance for wildfire occurrence in British Columbia. CatBoost reached the highest accuracy (0.92), precision (0.90), recall (0.87), and F1-score (0.88), followed closely by XGBoost with an F1-score of 0.85. Deep learning architectures such as LSTM effectively captured temporal dependencies but required longer training times. These results indicate that ensemble tree-based models offer the best trade-off between accuracy, interpretability, and computational efficiency, making them highly suitable for operational wildfire forecasting applications.

4.1. Computational Performance and Validation Strategy

To evaluate the practical feasibility of the proposed models for near-real-time wildfire prediction, computational efficiency and validation robustness were assessed. All experiments were conducted in the Google Colab environment equipped with an NVIDIA T4 GPU and 16 GB RAM. Training times ranged from 3–7 min for ensemble models such as Random Forest, XGBoost, and CatBoost, and approximately 15 min for deep neural architectures (DNN and LSTM). Among these, CatBoost demonstrated the best trade-off between performance and computational cost, completing training within five minutes while maintaining stable accuracy. Memory usage remained within 3–3.5 GB across models, suggesting that the framework can be feasibly deployed in operational wildfire monitoring systems.

Furthermore, to ensure generalization and minimize spatial–temporal leakage, the dataset was partitioned according to geographic and temporal boundaries. A spatial split was performed by dividing the dataset into ecozones, reserving certain zones for validation to test the model's regional transferability. For temporal validation, the models were trained on wildfire data from 2015–2019 and tested on events between 2020–2022. Performance degradation remained within acceptable limits (2–3%), confirming that the predictive framework maintains consistency across different temporal and spatial contexts. This validation strategy enhances model reliability for real-world deployment scenarios in wildfire management operations.

4.2. Baseline Comparison with Operational Indices

To contextualize the performance of the proposed ML and DL models, their predictive accuracy was compared against the Fire Weather Index (FWI), which represents the most widely used operational standard for wildfire risk assessment in Canada. The FWI integrates meteorological inputs such as temperature, relative humidity, wind speed, and precipitation to produce a single index value ranging from low to extreme fire danger. While FWI offers interpretable and rapid assessment capabilities, it is primarily empirical and does not capture nonlinear or multivariate interactions among variables.

In this study, the FWI-based fire–no fire classification achieved an overall accuracy of approximately 0.71 and an F1-score of 0.68 on the test dataset. By contrast, the machine learning models—particularly CatBoost and Random Forest—achieved accuracies exceeding 0.90 and F1-scores above 0.88. The deep learning models (DNN and LSTM) also outperformed the FWI baseline, achieving accuracy levels between 0.86 and 0.89. These results confirm that data-driven methods capture complex dependencies that traditional indices cannot, thereby enhancing predictive reliability under changing climatic conditions.

Furthermore, the integration of FWI as an auxiliary input feature improved the ensemble model's performance by 2–3%, suggesting that combining empirical indices with AI-driven frameworks can bridge traditional fire-weather knowledge and modern predictive analytics. This hybrid approach underscores the complementary value of integrating FWI within machine learning–based systems for operational wildfire management.

5. Conclusions and Future Works

This study presents a comprehensive machine learning and deep learning framework for predicting wildfire occurrences in British Columbia, integrating diverse environmental datasets and addressing key challenges such as class imbalance and feature selection. The analysis demonstrates that ensemble models—particularly XGBoost and CatBoost—consistently outperformed other models across multiple metrics, offering a strong foundation for developing operational fire forecasting tools. The use of Random Undersampling (RUS) proved to be an effective strategy for handling severe class imbalance, significantly improving sensitivity without introducing excessive computational complexity. Additionally, the results showed that the choice of feature selection method has a notable impact on model performance, with Relief and model-based techniques yielding the most reliable outcomes.

The findings of this study have important real-world implications. Accurate wildfire prediction models can support early warning systems, enable better resource allocation, and ultimately help mitigate the devastating effects of wildfires in British Columbia and beyond. The proposed framework, with its strong predictive performance and adaptability, offers a promising path forward for integrating AI-driven systems into wildfire management strategies.

Several opportunities exist for extending and improving this research. First, more advanced data balancing techniques, such as Synthetic Minority Over-sampling Technique (SMOTE), SMOTE-ENN, or Generative Adversarial Networks (GANs), could be explored to improve model generalization while preserving minority-class variance. Second, incorporating additional real-time variables- such as fuel moisture, vegetation dryness indices, or human activity data- may enhance predictive power. Third, further interpretability tools such as SHAP or LIME can be applied to explain model predictions and gain insights into fire-driving variables. Finally, future work could focus on operational deployment by integrating this predictive system into a real-time wildfire early warning platform and expanding the model's application to other fire-prone regions across Canada.

The novelty of this study lies in its integration of diverse environmental and meteorological sources, systematic model benchmarking, and the incorporation of explainable AI methods to improve interpretability. Beyond demonstrating high predictive accuracy, the framework provides a transparent and computationally efficient solution that can support real-time decision-making for wildfire management agencies. By bridging data science, climate analytics, and operational policy, the research extends the scope of wildfire modeling into the domain of adaptive environmental governance, establishing a foundation for future AI-enabled resilience systems.

Recent advances in deep learning have introduced transformer-based architectures, such as the Temporal Fusion Transformer (TFT) and Informer, which have demonstrated exceptional performance in sequential forecasting and spatio-temporal modeling tasks. Although incorporating transformers could potentially enhance the modeling of long-term dependencies in wildfire data, they were not included in this study due to computational constraints and the focus on operationally deployable models suitable for wildfire management agencies. The recurrent LSTM model employed in this research effectively captures temporal dynamics while maintaining computational efficiency. Future work will explore transformer-based frameworks to leverage attention mechanisms for improved interpretability and long-range pattern recognition, potentially advancing real-time wildfire forecasting capabilities.

Major Takeaways

- Ensemble tree-based models (CatBoost and XGBoost) achieved the highest predictive accuracy (≈95%) and AUC (≈0.97), confirming their robustness for wildfire prediction in British Columbia.
- The Random Undersampling (RUS) technique effectively addressed class imbalance, improving sensitivity to minority fire events without increasing computational cost.
- Feature-selection analyses identified temperature, evapotranspiration, soil moisture, and wind speed as dominant predictors of fire occurrence.
- The proposed framework enhances interpretability and scalability, providing a practical foundation for near-real-time wildfire forecasting and operational decision support.

Author Contributions: Conceptualization, M.N. and K.P.; methodology, M.N. and K.P.; software, M.N.; validation, M.N. and K.P.; formal analysis, M.N.; investigation, M.N.; resources, M.N. and K.P.; data curation, M.N.; writing—original draft preparation, M.N.; writing—review and editing, M.N. and K.P.; visualization, M.N.; supervision, K.P.; project administration, K.P.; funding acquisition, K.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The integrated wildfire dataset developed in this study was constructed by combining publicly available data from ERA5 reanalysis (https://cds.climate.copernicus.eu, accessed on 11 May 2025), the Canadian Wildland Fire Information System (https://cwfis.cfs.nrcan.

gc.ca, accessed on 11 May 2025), NASA FIRMS (https://firms.modaps.eosdis.nasa.gov/, accessed on 11 May 2025), and the BC Wildfire Service (https://www2.gov.bc.ca/gov/content/safety/wildfire-status, accessed on 11 May 2025). The processed and integrated dataset generated by the authors is not publicly available at this stage due to its large size and ongoing institutional review, but it can be obtained from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Flannigan, M.; Stocks, B.; Wotton, M. Impacts of climate change on fire activity and fire management in the circumboreal forest. *Glob. Change Biol.* **2009**, *15*, 549–560. [CrossRef]
2. Wotton, B.M.; Nock, C.A.; Flannigan, M.D. Forest fire occurrence and climate change in Canada. *Int. J. Wildland Fire* **2017**, *26*, 985–999. [CrossRef]
3. Hanes, C.C.; Wang, X.; Parisien, K.B.; Little, M.A.; Flannigan, M.D. Fire-regime changes in Canada over the last half century. *Can. J. For. Res.* **2019**, *49*, 256–269. [CrossRef]
4. Gillett, N.P.; Weaver, A.J.; Zwiers, F.W.; Flannigan, M.D. Detecting the effect of climate change on Canadian forest fires. *Geophys. Res. Lett.* **2004**, *31*, L18211. [CrossRef]
5. Hirsch, K.; Fuglem, P.; Kafka, V. *Wildfire Management in Canada: Review and Perspectives*; Canadian Forest Service: Ottawa, ON, Canada, 2021.
6. Government of Canada. Wildfire Causes and Risks. 2021. Available online: https://www.nrcan.gc.ca/climate-change/impacts-adaptations/climate-change-impacts/wildfires/10771 (accessed on 11 April 2025).
7. NASA FIRMS. Fire Information for Resource Management System. 2021. Available online: https://firms.modaps.eosdis.nasa.gov/ (accessed on 1 September 2025).
8. Government of British Columbia. Wildfire Management Branch. 2022. Available online: https://www2.gov.bc.ca/gov/content/safety/wildfire-status (accessed on 11 April 2025).
9. Parisien, M.A.; Kafka, V.; Hirsch, K.G.; Todd, J.B.; Lavoie, S.G.; Maczek, P.D. *Mapping Wildfire Susceptibility with the Burn-P3 Simulation Model*; Information Report NOR-X-405; Canadian Forest Service: Edmonton, AB, Canada, 2005.
10. BC Wildfire Service. Wildfire Statistics and Information. 2022. Available online: https://www2.gov.bc.ca/ (accessed on 11 May 2025).
11. FireSmart Canada. FireSmart Begins at Home Manual. 2021. Available online: https://www.firesmartcanada.ca (accessed on 11 May 2025).
12. Martell, D.L.; Otukol, S.; Stocks, B.J. A logistic model for predicting daily people-caused forest fire occurrence in Ontario. *Can. J. For. Res.* **1989**, *19*, 256–263. [CrossRef]
13. Finney, M.A. *FARSITE: Fire Area Simulator–Model Development and Evaluation*; RMRS-RP-4 Revised; USDA Forest Service: Ogden, UT, USA, 2004.
14. Littell, J.S.; McKenzie, D.; Peterson, D.L.; Westerling, A.L. Climate and wildfire area burned in western U.S. ecoprovinces, 1916–2003. *Ecol. Appl.* **2009**, *19*, 1003–1021. [CrossRef] [PubMed]
15. Szpakowski, D.M.; Jensen, J.L.R. A Review of the Applications of Remote Sensing in Fire Ecology. *Remote Sens.* **2019**, *11*, 2638. [CrossRef]
16. Giglio, L.; Randerson, J.T.; van der Werf, G.R. Analysis of daily, monthly, and annual burned area using the fourth-generation global fire emissions database (GFED4). *J. Geophys. Res. Biogeosci.* **2013**, *118*, 317–328. [CrossRef]
17. Jain, P.; Coogan, S.C.P.; Subramanian, S.G.; Crowley, M.; Taylor, S.W.; Flannigan, M.D. A review of machine learning applications in wildfire science and management. *Environ. Rev.* **2020**, *28*, 478–505. [CrossRef]
18. Rodriguez-Galiano, V.F.; Ghimire, B.; Rogan, J.; Chica-Olmo, M.; Rigol-Sanchez, J.P. An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS J. Photogramm. Remote Sens.* **2012**, *67*, 93–104. [CrossRef]
19. Zhang, J.; Li, L.; Zhu, Q.; Zhou, J. Wildfire susceptibility mapping using deep learning in Canada. *Remote Sens. Lett.* **2019**, *10*, 636–645.
20. Omar, N.; Al-zebari, A.; Sengur, A. Deep learning approach to predict forest fires using meteorological measurements. In Proceedings of the 2021 2nd International Informatics and Software Engineering Conference (IISEC), Ankara, Turkey, 16–17 December 2021; Department of Information Technology, Duhok Polytechnic University: Duhok, Iraq, 2023.
21. Hong, H.; Tsangaratos, P.; Ilia, I.K.; Liu, J.; Zhu, A.-X.; Xu, C. Applying genetic algorithms to set the optimal combination of forest fire related variables and model forest fire susceptibility based on data mining models: The case of Dayu County, China. *Sci. Total Environ.* **2018**, *630*, 1044–1056. [CrossRef] [PubMed]
22. Tavakoli, F.; Naik, K.; Zaman, M.; Purcell, R.; Sampalli, S.; Mutakabbir, A.; Lung, C.-H.; Ravichandran, T. Big Data Synthesis and Class Imbalance Rectification for Enhanced Forest Fire Classification Modeling. In Proceedings of the 16th International

Conference on Agents and Artificial Intelligence (ICAART), Rome, Italy, 24–26 February 2024; SciTePress: Barcelona, Spain, 2024; Volume 2, pp. 264–275. [CrossRef]
23. Elsarrar, O.; Darrah, M.; Devine, R. Analysis of forest fire data using neural network rule extraction with human understandable rules. In Proceedings of the 2019 18th IEEE International Conference on Machine Learning and Applications (ICMLA), Boca Raton, FL, USA, 16–19 December 2019; West Virginia University: Morgantown, WV, USA, 2020.
24. Bayat, G.; Yildiz, K. Comparison of the machine learning methods to predict wildfire areas. *Turk. J. Sci. Technol.* **2022**, *17*, 241–250. [CrossRef]
25. Sharma, U.; Shaw, S.; Kumari, K.S.; Shailendra, A.; Bengani, C.; Ramesh, S. Forest fire prediction using supervised machine learning algorithms. In Proceedings of the International Conference on Recent Trends in Data Science and Its Applications (ICRTDA 2023), Kattankulathur, India, 30–31 March 2023; SRM Institute of Science and Technology: Kattankulathur, Tamil Nadu, India, 2022.
26. Li, Y.; Feng, Z.; Chen, S.; Zhao, Z.; Wang, F. Application of the artificial neural network and support vector machines in forest fire prediction in the Guangxi Autonomous Region, China. *Discret. Dyn. Nat. Soc.* **2020**, *2020*, 5612650. [CrossRef]
27. Olawade, D.B.; Bamisile, O.; Adedeji, K. Artificial intelligence in environmental monitoring: A comprehensive review. *Environ. Adv.* **2024**, *18*, 100627.
28. Camps-Valls, G. Fernández-Torres, M.; Cohrs, K.-H.; Höhl, A.; Castelletti, A.; Pacal, A.; Robin, C.; Martinuzzi, F.; Papoutsis, I.; Prapas, I.; et al. Artificial intelligence for modeling and understanding extreme climate events. *Nat. Commun.* **2025**, *16*, 56573.
29. Forouheshfar, Y.; Ayadi, R.; Moghadas, O. Enhancing system resilience to climate change through artificial intelligence: A systematic literature review. *Front. Clim.* **2025**, *7*, 1585331. [CrossRef]
30. Wu, X.; Chen, H.; Zhang, Y. Machine learning-based prediction of resilience in green supply chain systems. *Systems* **2025**, *13*, 615. [CrossRef]
31. Canadian Wildland Fire Information System (CWFIS). Natural Resources Canada. 2022. Available online: https://cwfis.cfs.nrcan.gc.ca/ (accessed on 11 April 2025).
32. Environment Canada. Historical Climate Data. 2022. Available online: https://climate.weather.gc.ca/ (accessed on 11 April 2025).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

fire

MDPI

Article

TCSN-YOLO: A Small-Target Object Detection Method for Fire Smoke

Cao Yang [1,*], Zhou Jun [2], Wen Hongyuan [1] and Wang Gang [1]

[1] School of Intelligent Manufacturing, Taizhou Institute of Science and Technology, Nanjing University of Science Technology, Taizhou 225300, China; 22009025@nustti.edu.cn (W.H.); wanggang@njust.edu.cn (W.G.)
[2] School of Electronic and Electrical Engineering, Taizhou Institute of Science and Technology, Nanjing University of Science Technology, Taizhou 225300, China
* Correspondence: caoyang@njust.edu.cn

Abstract

Forest fires continue to pose a significant threat to public and personal safety. Detecting smoke in its early stages or when it is distant from the camera is challenging because it appears in only a small region of the captured images. This paper proposes a small-scale smoke detection algorithm called TCSN-YOLO to address these challenges. First, it introduces a novel feature fusion module called trident fusion (TF), which is innovatively designed and incorporated into the neck of the model. TF significantly enhances small target smoke recognition. Additionally, to obtain global contextual information with high computational efficiency, we propose a Cross Attention Mechanism (CAM). CAM captures diverse smoke features by assigning attention weights in both horizontal and vertical directions. Furthermore, we suggest using SoftPool to preserve more detailed information in the feature map. Normalized Wasserstein Distance (NWD) metric be embedded into the loss function of our detector to distinguish positive and negative samples under the same threshold. Finally, we evaluate the proposed model using AI For Humankind dataset and FlgLib dataset.The experimental results demonstrate that our method achieves 37.1% APs, 90.3% AP50, 40.4% AP50:95, 45.34 M Params and 170.5 G FLOPs.

Keywords: small target; trident fusion; cross attention mechanism; SoftPool; NWD-loss

1. Introduction

Forest resources are vital strategic assets for humanity, serving essential economic, ecological, and social functions. Forest fires not only pollute water and air but also pose significant threats to human lives and property. Therefore, it is crucial to accurately detect forest fires and ensure the safety of human life and property [1]. During the early stages of forest fires, the smoke area is small and the spread rate is slow, making it the optimal time to extinguish the fire [2,3]. However, in the field of computer vision, it is a challenge to quickly and accurately identify targets at a long distance and small size, especially forest fires with a complex background. Take the definition of Common Objects in Context (COCO) object as an example, a small object refers to less than 32 × 32 pixels (medium object refers to 32 × 32–96 × 96, and large object refers to more than 96 × 96) [4]. How to extract and identify these small target features with less calculation is the focus of this study.

The traditional vision-based detection approach focuses on the artificial design of the smoke detection features [5]. Prema et al. [6] splits the smoke region based on the

color of the local YUV color, and extract the static and dynamic texture features, such as uniformity, energy, correlation, and contrast. Filonenko et al. [7] utilized color and shape features for smoke detection. However, the color of the smoke often depends on the burning material, and its shape is easily influenced by wind and airflow. Alamgir et al. [8] used local binary patterns combined with co-occurrence texture features to characterize smoke. Li et al. [9] proposed a flame detection framework that automatically determines the Gaussian distribution of flame color using Dirichlet process estimation. Krstinić et al. [10] use HSI color space transformations to achieve pixel-level smoke segmentation.Whether it is based on flame color or smoke morphology and texture, these methods rely on the researchers' choices and designs, and cannot fully represent the characteristics of smoke or the connections between these characteristics. Consequently, traditional visual inspection methods have poor generalization, and interference from the environment significantly impacts detection results.

As the second visual detection method, deep learning is utilized to extract smoke features [11]. Wang et al. [12] proposed a lightweight smoke detection model ESmokeNet, which includes an innovative EdgeDet head that can capture the edge information of smoke to improve the accuracy of the final decision. At the same time, a context embedding module is added to the backbone network of the model to enhance the extraction and fusion capabilities of multi-scale features. Ba et al. [13] enhanced the feature representation capability by a residual attention module (RA), which integrates the residual module with the spatial and channel attention mechanisms. Hashemzadeh et al. [14] propose a method that combines the visual feature extraction capabilities of convolutional neural networks (CNNs) with an efficient spatiotemporal feature calculation method to achieve accurate and real-time detection of smoke in video sequences. Majid et al. [15] propose a custom fire detection framework based on transfer learning, which uses Gradient Weighted Class Activation Mapping (Grad-CAM) to generate a coarse localization map that highlights important regions in the image for predicting concepts. Khan et al. [16] introduced a model termed DeepSmoke based on EfficientNet and DeepSmoke 3+. Hosseini et al. [17] propose a method for detecting smoke and flames in video frames, called the UFS network. The network employs a voting-based decision-making scheme, which can reduce false positives that occur with CNNs and improve the reliability of the system as much as possible. Zhu et al. [18] proposed the small object detection algorithm TPH-YOLOv5 for drone-captured images. This algorithm adds a small object detection head based on the YOLOv5 algorithm, which can detect small objects on feature layers with richer pixels. At the same time, transformer encoder blocks are also integrated in the backbone and neck sections, increasing the ability to capture global information. To address the computational overhead caused by the additional prediction heads in the TPH-YOLOv5 architecture, Zhao et al. [19] removed the tiny object prediction head and added a cross-layer asymmetric transformer (CA Trans) to reduce the amount of computation while retaining the performance of small object detection. Amjad et al. [20] propose a Dynamic Fire and Smoke Detection Model (DFDM), which is an optimized YOLOv7-tiny architecture that incorporates an asymptotic feature pyramid network (AFPN) to bridge the semantic gap and introduces a cross-layer dual attention (CDA) mechanism to improve the detection capability of key fire and smoke features. Bahhar et al. [21] proposed a two-stage YOLO model. The first stage uses YOLOv5s to detect smoke, and the second stage uses the standard YOLOv5 with a voting ensemble CNN architecture to give the final decision. To effectively reduce the false positive rate during smoke detection, Kwak et al. [22] use the dark channel prior algorithm to filter out interfering information in the image background, such as fog. We also use the optical flow method to extract the dynamic features of the smoke. Jeong et al. [23] A combined model based on the YOLO detector and the long short-term memory (LSTM)

classifier with a teacher-student framework is proposed to improve detection performance by reflecting the spatiotemporal characteristics of smoke. The above-mentioned models have good performance in the field of target detection. We believe that after multiple convolutions, the features of small objects are likely to disappear. Furthermore, convolution operations cannot resolve the long-range dependency of feature information, making it difficult to improve detection accuracy.

2. Related Work

The attention mechanism can improve the performance of target detection models by preserving global contextual features significantly and extracting more differentiated information. Woo et al. [24] proposed a lightweight Convolutional Block Attention Module (CBAM) to adaptively refine feature maps by emphasizing informative features along both channel and spatial dimensions, thereby helping the model focus on key information. CBAM can be easily integrated into Convolutional Neural Network (CNN) architectures to highlight relevant features and improve their representation. DETR [25], as a novel Transformer architecture designed for object detection tasks, which uses a self-attention mechanism and positional embedding to capture global contextual information and understand the positional relationships of objects in images. Liu et al. [26] proposed a Swin transformer to captures longrange dependencies by self-attention. Safarov et al. [27] proposed a novel feature extraction structure, which replaces the CSPDarknet53 in the YOLOv5 backbone network with the vision transformer (ViT), and uses the self-attention mechanism of ViT to obtain the long-range dependency of the target. In particular, the detection results of the model are significantly improved in the presence of interference. However, algorithms based on the self-attention mechanism have problems such as long training time.

Feature fusion prevents information loss. Rohra et al. [28] proposed a multi-scale feature fusion network (MSFFNet), which extract sufficient semantic features to understand targets in both sparse and highly crowded scenes. This method employs adaptive fusion to integrate most encoded features, rather than extracting features separately. Therefore, it can expand the receptive field size and reduce computational cost. Singh [29] proposed a dual-stream feature fusion (BFF) network for target detection in hazy environments. The network's multi-level feature fusion unit can fuse appropriate features between hazy images and mixed inputs, thus achieving excellent performance in hazy environments. Ghiasi et al. [30] proposed an feature fusion network based on neural architecture search (NAS-FPN), which automatically generates high-performance, scalable feature pyramid structures through neural architecture search, solving the suboptimal problem of manually designed FPNs and achieving synergistic optimization of accuracy and efficiency in multi-scale target detection tasks. Achinek et al. [31] introduce a dense attention mechanism in the Feature Network (FPN) layer, which can propagate semantic information and reduce the learning of redundant feature maps.

A good regression loss function will significantly improve the model detection performance. In the early target detection model based on anchor box, IoU is usually used to measure the overlap between the anchor box and the GT box. It has fast convergence speed and scale invariance, but it cannot accurately reflect the coincidence degree between them. DIoU adds the distance information of the center points of two boxes compared with IoU. The author indicates that a good BBR should consider three indicators, namely, overlap, center distance and aspect ratio, while DIoU only considers the first two, which solves the problems of divergence in training when IoU is used as a loss function. CIoU contains all three indexes, and combines the influence factors of aspect ratio into the penalty term, so it has a faster convergence speed. Focaler-IoU reformulates the IoU loss function through

a linear interval mapping methodology, establishing a dynamic adaptation mechanism that differentiates treatment for regression samples based on their difficulty levels, thereby optimizing the model's focus on critical learning objectives [32]. This method performs particularly well in scenes with small targets.

In order to improve the shortcomings of current research and improve the detection accuracy of smoke, this study designs an efficient feature extraction and feature fusion model for detecting small-target forest fire smoke. In the neck part of the network structure, the algorithm innovatively introduces shallow feature maps from the backbone to obtain richer small target information. Furthermore, this study investigates the method of focusing on the texture feature information of fire smoke that uses sparse attention to compute long-range dependencies of contextual information efficiently. In order to preserve more detailed feature information during feature extraction, we replace MaxPool pooling in the SPPF layer with SoftPool pooling. Finally, we use the normalized Gaussian Wasserstein distance as the basis for calculating the bounding box regression loss, which can solve the problem that IoU loss is very sensitive to the position deviation of tiny objects.

The main contributions of this work are as follows:

1. We propose a trident fusion (TF) module to integrate deep features, corresponding features, and shallow features to obtain more shallow small target information.
2. We propose a Cross Attention Mechanism (CAM) to capture smoke texture features in both horizontal and vertical dimensions.
3. We replace MaxPool with SoftPool in the SPPF layer of YOLOv8 to minimize information loss during the pooling process while preserving the functionality of the pooling layer.
4. Instead of the original CIoU metric, NWD is embedded into the bounding box regression loss function of the detector model.
5. The performance of this algorithm was analyzed and compared with other mainstream object detection algorithms.

The rest of this article is organized as follows: Section 2 discusses existing research on smoke recognition, and Section 3 introduces the TCSN-YOLO detection algorithm for small smoke targets. Section 4 presents test verification and analysis. Finally, Section 5 summarizes the findings and elaborates on future research directions.

3. Materials and Methods

The overall network structure of the proposed small smoke target detection model named TCSN-YOLO is shown in Figure 1. The proposed Trident Fusion (TF) module fuses shallow and deep feature maps and integrates Soft Pooling to enhance the efficiency of deep feature extraction and minimize information loss during downsampling. In addition, the Cross Attention Mechanism (CAM) is connected after the soft SPPF, which can improve the network's attention to the region of interest from both horizontal and vertical directions. Finally, NWD-based Regression Loss applied to our detector avoids minor location deviation flip the anchor label.which enhance positive/negative sample differentiation,and reduces the convergence difficulty of the network.

3.1. Trident Fusion

We propose the Trident Fusion (TF) module, a novel three-dimensional feature map fusion mechanism that enhances spatial and semantic information for the detection head. As illustrated in Figure 2, the deep feature layer P5 and the middle feature layer P4 in the model backbone network serve as one input branch of TF module, the shallow feature layer P3 serves as another input branch. The feature layer P2 shallower than the feature layer P3 has richer pixel information, which serves as the third input. The TF

module adjusts the above three input feature maps to the same size and concatenates them together. Compared with the original feature fusion module of the baseline network, the proposed module preserves rich low-level features, thereby enhancing small-target detection performance. The shallow feature layer P2 has a size of 160 × 160 × 128. After depthwise convolution using DSC (kernel size = 3, s = 2, padding = 1), the size is reduced to 80 × 80 × 128. Then, through point-by-point convolution, information from different channels is linearly combined to form the feature layer P2′. The upsample module in the Trident Fusion module uses nearest neighbor interpolation to upscale the small input feature map (40 × 40 × 512) by a factor of 2 with minimal computational effort. Notably, the feature map P2 retains substantial positional information, which is critical for localizing small objects due to its high spatial resolution. However, the rich background information contains complex textural characteristics that may interfere with target detection results. Therefore, in TF module, a depthwise separable convolution (DSC) [33] is deployed to filter out irrelevant information related to the detected object.

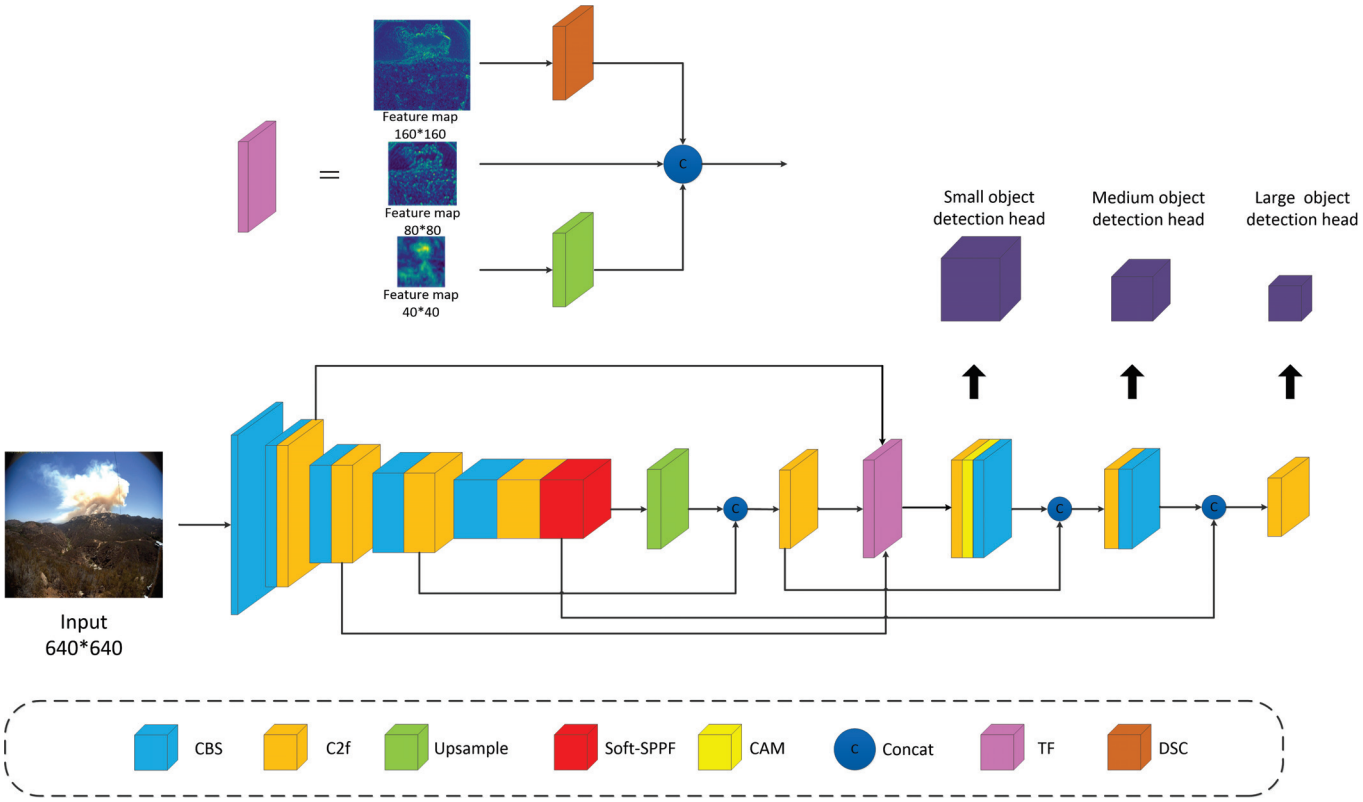


Figure 1. TCSN-YOLO network architecture.

DSC comprises two stages: a depthwise convolution that operates on single-channel feature maps and a pointwise convolution that operates across feature map channels. When using DSC to perform channel convolution and point convolution on a 160 × 160 × C_{in} feature map, the number of parameters is 3 × 3 × C_{in} + 3 × C_{out}, while using a traditional convolution kernel, the number of parameters is 3 × 3 × C_{in} × C_{out}, Where C_{in} refers to the number of input feature map channels, and C_{out} refers to the number of output feature map channels. DSC has relatively fewer parameters compared to traditional convolutional layers with kernels of equivalent size. This approach not only effectively extends the receptive field but also minimizes parameter growth.

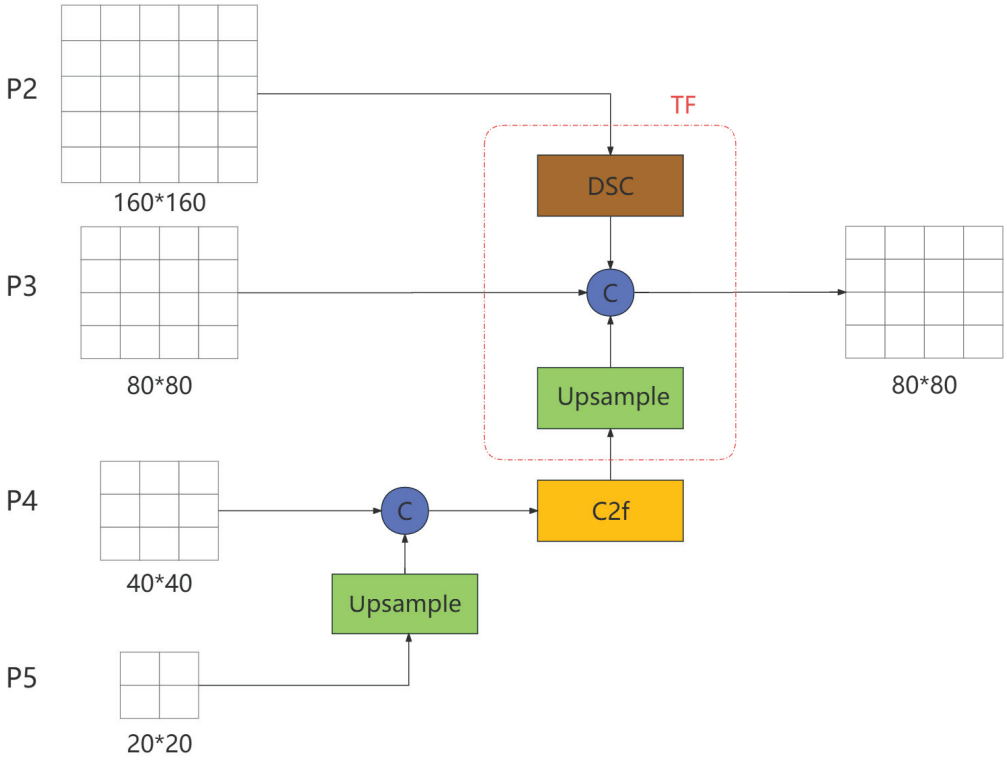


Figure 2. Trident Fusion module structure.

3.2. Cross Attention Module

We propose a Cross Attention Module (CAM) to capture contextual information with a minor computation increment for smoke detecting. The network structure is illustrated in Figure 3. It consists of three conv, two Self-Attention, an addition operation, and a Residual Connection. Since Self-Attention shows remarkable performance in object detection, we adopt Self-Attention as the Main Operations to extract horizontal and vertical directions features. which is input of addition operation. The residual connection facilitates cross-layer feature integration through additive operations, where shallow-level features are directly combined with their corresponding deep representations. This architectural design not only promotes hierarchical feature reuse but also mitigates gradient vanishing issues, thereby significantly enhancing the model's generalization capacity by preserving information propagation pathways throughout network depth.

The feature map X with spatial dimensions H × W × C is processed by the cross-attention module to yield the output feature map X′. Specifically, three convolutional layers with 1 × 1 filters are applied on feature map X to generate three feature maps Q, K and V, respectively, where Q, K $\in \mathbb{R}^{c' \times h \times w}$, V $\in \mathbb{R}^{c \times h \times w}$. c′ is 1/8 of c, with c being the original channel count in X. Then we adjust the dimensions of matrices Q and K to obtain horizontal directions feature maps Q_h, K_h and V_h, respectively. We further obtain the correlation degree of all pixels in horizontal directions, which is named Ah via matrix multiplication. The matrix multiplication operation is then defined as follows:

$$A_h = softmax(Q_h K_h) \tag{1}$$

where parameter matrices $Q_h \in \mathbb{R}^{w \times h \times c'}$, $K_h \in \mathbb{R}^{w \times c' \times h}$.

Vh based on V multiply with Ah to obtain horizontal directions feature matrix X'_h. The calculation process of attention in the vertical direction is similar to that in the horizontal direction, and the final feature matrix named $X'_w \in \mathbb{R}^{c \times h \times w}$.

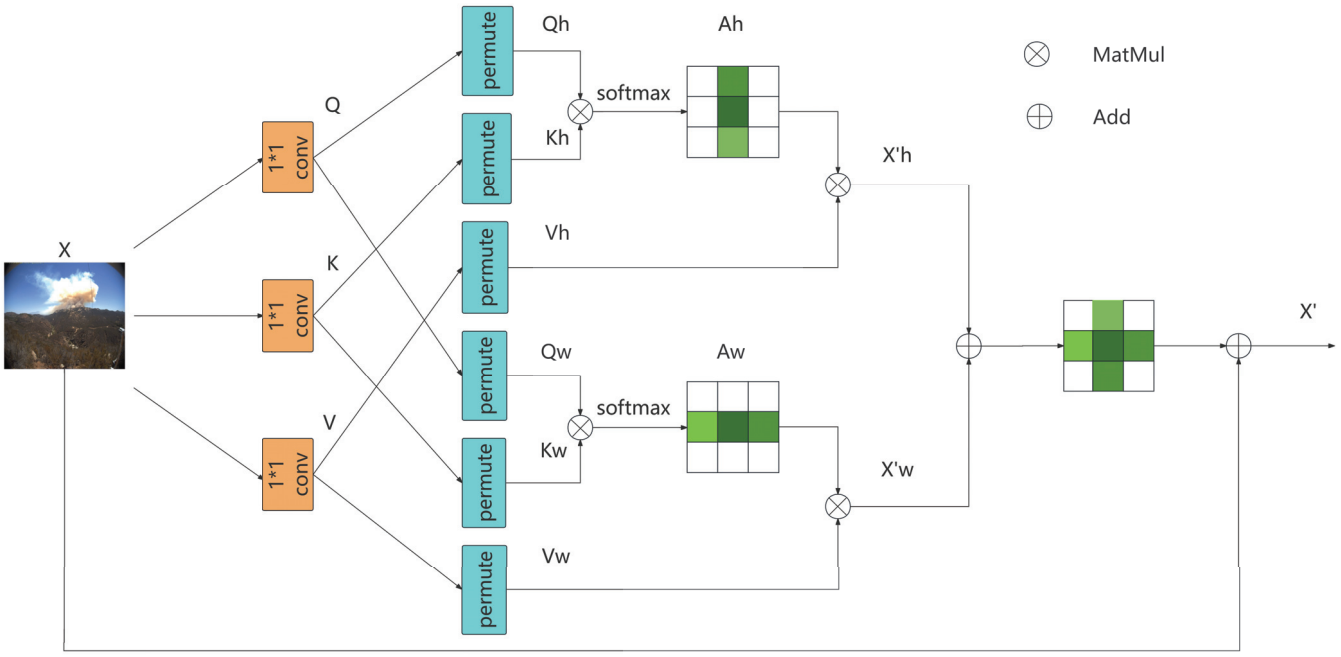


Figure 3. Cross attention module structure.

$$X'_h = A_h \times V_h \tag{2}$$

$$X'_w = A_w \times V_w \tag{3}$$

where parameter matrices $A_h \in \mathbb{R}^{w\times h\times h}$, $V_h \in \mathbb{R}^{w\times c\times h}$, $A_w \in \mathbb{R}^{h\times w\times w}$, $V_w \in \mathbb{R}^{h\times c\times w}$.

Finally, the feature matrix X'_h and X'_w are fused by using addition operation. It is follow by a shortcut connection to facilitate the transmission of gradient information. The operation is then defined as follows:

$$X' = (X'_h + X'_w) + X \tag{4}$$

Despite CAM capturing contextual information in both horizontal and vertical directions, it remains insufficient for comprehensively capturing long-range dependencies across all pixels. To tackle this problem, We connect two CAM modules in series. The first CAM takes the feature map X as its input, and the output is a feature map X' with the horizontal and vertical context information in the feature map X. The second CAM then takes X' as its input and outputs a feature map X'', where X'' has the same shape as X and contains the global context information in the feature map X.

The complexity of classic self-attention in time and space is $O((h \times w) \times (h \times w))$. By serially stacking two CAMs, it reduces the complexity in time to $O(((h \times w) \times (h + w)) \times 2)$, and CAM accelerates the speed of model training and inference greatly.

3.3. Soft-SPPF Module

Pooling reduces computational load by downsampling input feature maps while preserving salient features. Most architects use maximum or average pooling, but if maximum pooling is selected, most feature information will be lost. In the latter case, if average pooling is selected, the intensity of feature information in the whole area will be averaged, resulting in the weakening of local intensity. This paper adopts the SoftPool method, which balances global and local information by giving high weight to important

features. SoftPool is differentiable, allowing the minimum gradient value to be updated during backpropagation.

SoftPool is an softmax weighted activation subsampling method. The feature weights are derived from the following formula:

$$w_i = \frac{e^{a_i}}{\sum_{j \in R} e^{a_j}} \tag{5}$$

In the formula, the weight w_i is equal to the natural exponent of the activation value a_i divided by the sum of the natural exponent of all the activation values. Activation values a_i play a prominent role in the calculation method of weights, higher activation values will become more and more important, SoftPool is a more balanced method, rather than simply choosing the maximum value. Sum all weighted activation values to obtain SoftPool's output value:

$$\widetilde{a} = \sum_{i \in R} w_i \times a_i \tag{6}$$

We replace the maximum pool in the original SPPF with SoftPool, and name this structure the Soft-SPPF module. The structure of the Soft-SPPF module is illustrated in Figure 4. Initially, input feature maps pass through a convolutional layer for feature compression. Subsequently, three sequential SoftPool operations with a 5×5 kernel size are applied to generate multi-scale feature maps. Finally, these feature maps with different scales are spliced in the channel dimension to form a feature vector containing multi-scale information by the "Concat" operation and a convolution layer. During this process, the Soft-SPPF module compresses the number of channels of the input feature map from 512 to 256. After multi-scale feature fusion, the number of channels in the output feature map is restored to 512, significantly reducing the amount of computation.

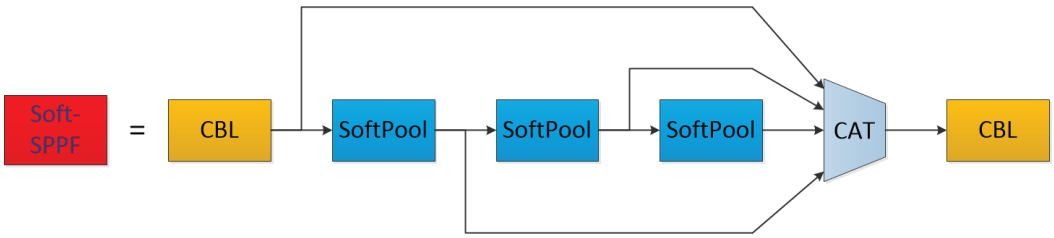


Figure 4. Soft-SPPF module structure.

3.4. Normalized Gaussian Wasserstein Distance Loss

The Normalized Wasserstein Distance (NWD) is an evaluation metric specifically designed for small object detection scenarios [34]. By modeling bounding boxes as Gaussian distributions and computing the Wasserstein distance between them, NWD offers a more nuanced, differentiable, and robust alternative to IoU, as defined by Equation (7). Crucially, NWD effectively measures similarity regardless of bounding box overlap, overcoming a limitation of traditional intersection-based metrics. Furthermore, its inherent insensitivity to target scale renders NWD particularly suitable for assessing the similarity of small targets.

$$W_2^2(\mathcal{N}_a, \mathcal{N}_b) = \|(Q_a, Q_b)\|_2^2 \tag{7}$$

$$Q_a = \left[cx_a, cy_a, \frac{w_a}{2}, \frac{h_a}{2}\right]^{\mathrm{T}} \tag{8}$$

$$Q_b = \left[cx_b, cy_b, \frac{w_b}{2}, \frac{h_b}{2}\right]^{\mathrm{T}} \tag{9}$$

$W_2^2(\mathcal{N}_a, \mathcal{N}_b)$ denotes a distance metric between two Gaussian distributions, Where $\mathcal{N}_a, \mathcal{N}_b$ are modeled via bounding boxes Q_a = (cx_a, cy_a, $w_a/2$, $h_a/2$) and Q_b = (cx_b, cy_b, $w_b/2$, $h_b/2$).

$$NWD(\mathcal{N}_a, \mathcal{N}_b) = \exp\left(-\frac{\sqrt{W_2^2(\mathcal{N}_a, \mathcal{N}_b)}}{C}\right) \tag{10}$$

NWD is derived by applying exponential normalization to the Wasserstein distance metric, enabling its direct application as a similarity measure, where the constant C serves as a scale normalization parameter dependent on the dataset. Based on the average pixel value occupied by small smoke targets in the AI For Humankind dataset used for training, we choose 18 as the value of the normalization constant C.

$$\mathcal{L}_{NWD} = 1 - NWD(\mathcal{N}_a, \mathcal{N}_b) \tag{11}$$

We integrate NWD into the regression loss function to replace CIoU loss, which allows the model to achieve scale invariance while the performance of measuring the similarity between bounding boxes could still be maintained.

4. Results

4.1. Dataset

This paper utilizes the dataset, which is an open dataset collected and organized by the non-governmental organization AI For Humankind, to investigate the deployment of object detection models for early wildfire smoke identification [35]. This dataset has 2192 annotated images captured using HPWREN cameras with 640 × 480 pixel resolution. To enhance model generalization during training, data augmentation is performed probabilistically. The image is mirrored with a probability of 20%. The image is rotated by 10° with a probability of 20%. The image is scaled by a factor of 0.9 with a probability of 20%. The image brightness is adjusted by a value of 0.4 with a probability of 20%. The image is translated by a value of 0.1 with a probability of 20%, especially increasing the number of images with white clouds, strong light, and fog as backgrounds to verify the effectiveness of the model under environmental interference. Consequently, the expanded dataset comprised 12,620 images. The AI For Humankind dataset is randomly divided by 8:1:1. Some representative images are shown in Figure 5.

4.2. Implementation Details

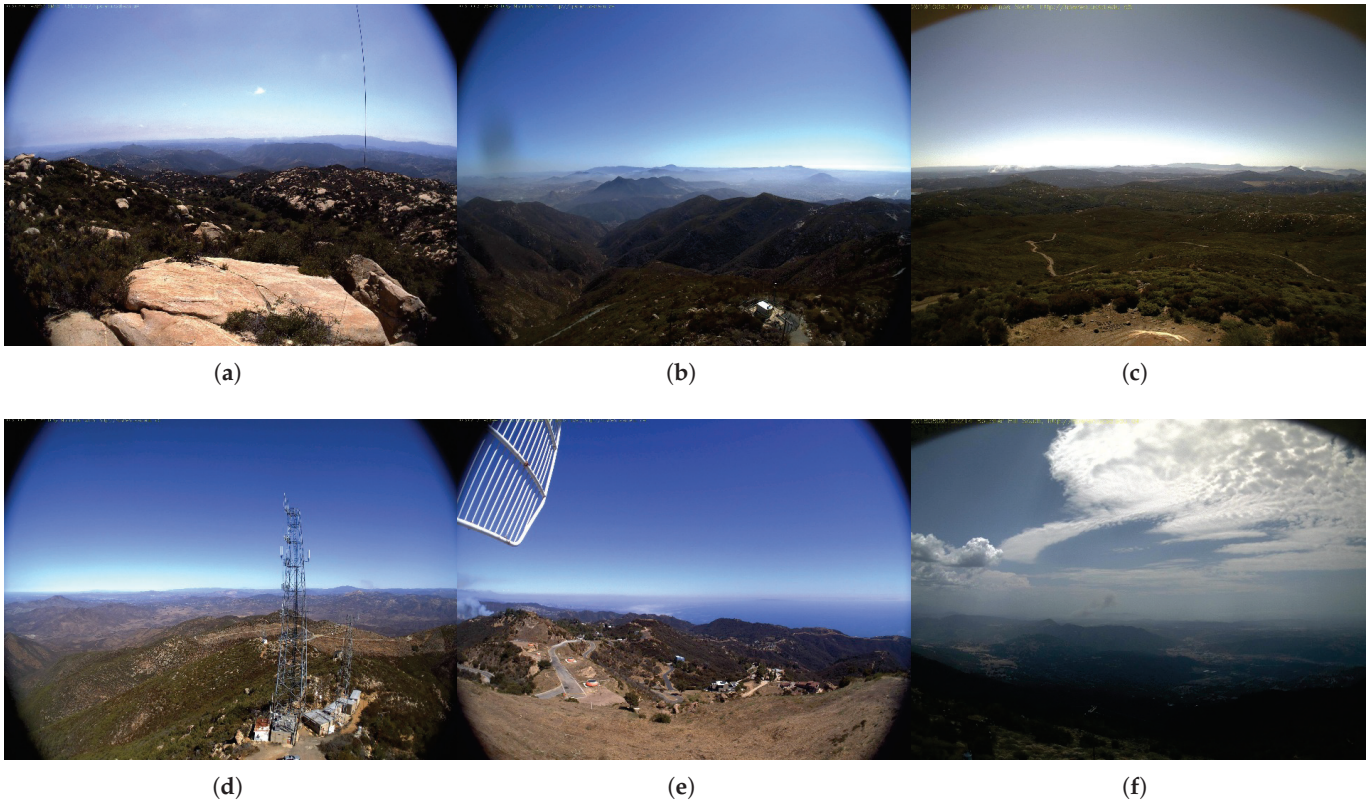
To evaluate and verify the performance of TCSN-YOLO, we conducted several experiments. The hardware and software environments were consistent across all tests. Detailed environmental information is provided in Table 1.

Table 1. Hardware and software environment.

Hardware environment	CPU	Intel Xeon Gold 6330
	RAM	64 G
	GPU	NVIDIA GeForce RTX 3090
	Video memory	16 G
Software environment	OS	Ubuntu 18.04
	language	Python 3.8
	framework	Pytorch 1.9.0

In the TCSN-YOLO, the activation function is an SiLU, which is optimized through Adaptive Motion Estimation (Adam) with a weight decay of 0.0005 and an initial learning

rate of 0.001. Non-Maximum Suppression (NMS) was implemented with an Intersection-over-Union (IoU) threshold of 0.5. For sample allocation, we employed a Task Alignment Assigner that dynamically allocates positive and negative samples by calculating the alignment score between the prediction box and the ground truth box.



(**a**) (**b**) (**c**)

(**d**) (**e**) (**f**)

Figure 5. Some representative images from AI For Humankind dataset. (**a**) Complex background, gray rocks. (**b**) Complex background, fog. (**c**) Complex background, strong light. (**d**) Simple background, small object. (**e**) Simple background, medium object. (**f**) Complex background, white clouds.

4.3. Experimental Model Evaluation Indicators

To visually evaluate the model's performance on small targets, we used APs (IoU = 0.50 for small smoke) as the main performance metric, and added AP50 (IoU > 0.5) and AP50:95 (0.95 > IoU > 0.5). In addition, considering that our proposed TCSN-YOLO application scenario is an embedded platform, the scale of the algorithm is limited by computing resources. In order to more accurately evaluate its performance, we also measure two pivotal computational complexity metrics: trainable parameter counts (Params) and floating-point operations (FLOPs).

1. Precision quantifies the proportion of correctly identified positive instances among all samples predicted as positive. The calculation formula is as follows:

$$P = \frac{TP}{TP + FP} \tag{12}$$

2. Recall measures the proportion of true positive instances correctly predicted by the model among all actual positive cases. The computational formula is expressed as follows:

$$R = \frac{TP}{TP + FN} \tag{13}$$

3. Average Precision.

$$AP = \int_0^1 P(R)dR \tag{14}$$

Average Precision (AP) quantifies the model's precision performance averaged across all recall levels for a specific category, typically computed as the area under the Precision–Recall (PR) curve.

4. Params. It quantifies the total parameter count of the model, while it does not directly influence the inference speed, it fundamentally determine memory requirements during inference.
5. FLOPs. It quantifies the computational complexity of model, serve as a key metric for evaluating model efficiency.

4.4. Ablation Analysis of Network Structures

YOLOv8 is used as the baseline for ablation experiments, which performs well on standard datasets such as Microsoft COCO. The YOLOv8L model achieves a mean average precision (mAP) of 52.9% when measured on COCO. This high accuracy stems from the C2f module adopted in its backbone which reduces parameters by 28% while enhancing feature reuse by shortening gradient paths and introducing cross-stage connections. Its neck architecture restructures the Feature Pyramid Network (FPN) and Path Aggregation Network (PAN), significantly improving hierarchical feature fusion across scales. These factors contribute to improved detection accuracy and efficiency.

To validate the effectiveness of each of the improvement strategies, we conduct a series of ablation experiments on the validation set of AI For Humankind dataset with different settings. As illustrated in Table 2, where ✓denotes strategy implementation, each enhancement strategy contributed to improved detection performance beyond the baseline model. Specifically, the integration of SoftPool, TF, CAM, and NWD loss yielded measurable gains, with all strategies collectively elevating key detection metrics.

Table 2. Results of ablation experiment.

Baseline	TF	CAM	SoftPool	NWD-Loss	AP_{50} (%)	$AP_{50:95}$ (%)	APs (%)	AP_M (%)	Params (M)	FLOPs (G)
✓					88.4	37.8	36.4	56.8	43.71	165.2
✓	✓				**89.5**	**38.7**	36.9	57.4	43.76	166.1
✓		✓			88.9	38.2	36.6	**58.2**	**44.18**	**167.4**
✓			✓		88.6	38.1	36.5	57.1	43.82	165.9
✓				✓	88.9	38.5	**37.0**	56.9	43.71	165.4
✓	✓	✓	✓	✓	90.3	40.4	37.1	61.6	45.34	170.5

The proposed Softpool enhances the accuracy of the network model in practical applications, improving AP50 by 0.2% and AP50:95 by 0.3%. The exponential weighting preserve the basic attributes of the input feature, while amplifying stronger feature activations.

The integration of the TF module into the feature fusion structure introduces shallow features rich in location information from the backbone network, thereby contributing to the clear details of the small targets. In addition, the TF module filters the background noise in the shallow features by using the DSC module. Baseline + TF module achieves 36.9% APs, which is 0.5% higher than that of Original YOLOv8 with similar params (43.76 M vs. 43.71 M) and slightly higher FLOPs (166.1 G vs. 165.2 G).

The CAM module extracts contextual features from the feature map along both the horizontal and vertical axes. It can obtain long-distance dependencies between pixels with a small increase in the amount of computation, which enable the model to balance computational efficiency and detection performance enhancement.

Integrating NWD into the regression loss function to replace CIoU loss ensures that the detection results of small targets will not be judged as negative samples due to slight deviations from the ground truth, thereby significantly improving small-target detection robustness. NWM-loss has significant improvements in AP50, AP50:95, and APs indicators, among which the APs indicator effect is better than other improvement strategies, proving that scale invariance plays a critical role in boosting small-target detection performance.

The proposed TCSN-YOLO model uses multiple improvement strategies. For small objects with complex environments, the proposed model shows excellent detection performance, which achieves higher confidence and reduces false positives and false negatives. Compared with the baseline model, the TCSN-YOLO model exhibits increases of 1.9% in AP50, 2.6% in AP50:95, and 0.7% in APs.

4.5. Comparison with Mainstream Methods

To evaluate the detection performance of TCSN-YOLO, we conducted comparative experiments with several mainstream object detection models, including TPH-YOLOv5, TPH-YOLOv5++, DETR and YOLOv8. Experiments were performed on the AI For Humankind dataset, and the results are summarized in Table 3. Comparative analysis reveals that the proposed TSCD-YOLO model achieved the highest detection precision among all mainstream models. Although the improved model had a slightly higher FLOPs compared to the TPH-YOLOv5++ and YOLOv8 model.

Table 3. Comparative performances of the proposed and mainstream models on AI For Mankind dataset.

Model	AP_{50} (%)	$AP_{50:95}$ (%)	AP_S (%)	AP_M (%)	Params (M)	FLOPs (G)
TPH-YOLOv5	87.5	37.5	35.1	58.0	45.36	260.1
TPH-YOLOv5++	87.7	38.8	36.1	59.5	**41.49**	**160.0**
DETR	78.8	30.0	29.6	55.4	60.28	253.6
YOLOv8	88.4	37.8	36.4	56.8	43.71	165.2
TCSN-YOLO (ours)	**90.3**	**40.4**	**37.1**	**61.6**	45.34	170.5

The experimental results can be explained as follows: Despite significantly improving small object detection performance, the additional prediction head in TPH-YOLOv5 also leads to large computational and memory overhead. Moreover, it is not scale-invariant and prone to false detection and missed detection. TPH-YOLOv5++ is an improvement on the TPH-YOLOv5 model. It eliminates the dedicated tiny object detection head. Instead, it employs the Cross-layer Asymmetric Transformer (CA-Trans) module to enhance small object feature expression, and models sparse cross-layer attention through the Sparse Local Attention (SLA) module, while reducing the amount of calculation, it maintains a high accuracy of small target detection. However, SLA only calculates the attention value between neighbors and lacks the process of capturing global contextual relationships. DETR adopts the Transformer codec architecture and uses its global attention mechanism to analyze the contextual information of the entire image. It shows significant advantages on large targets, but its performance on small targets still needs to be improved. Compared to the previously mentioned models, YOLOv8 has fewer parameters, lower FLOPs, and enhanced object detection performance. However, in a complex backgrounds, deep feature maps frequently contain insufficient semantic information to maintain robust detection performance for small objects. Through the efficient integration of complementary shallow feature maps, the detection model gains more comprehensive and effective representations, significantly enhancing small object detection under these challenging conditions. Consequently, TCSN-YOLO maintains high detection performance for small objects in complex

backgrounds with lower computational complexity and memory overhead. In contrast, competing methods exhibit degraded performance under similar conditions.

To further evaluate the detection performance of the models, we selected some images with complex backgrounds of fog, strong light and white clouds to compare the mainstream models and the TCSN-YOLO model. As shown in Table 4, the a column is the result without obvious interference information, the b column is the detection result under strong light illumination, the column c presents the detection result under the scene with white clouds in the picture background, and the column d presents the detection result under the scene with fog in the picture background. It show that the TCSN-YOLO model is little affected by background interference and has good robustness to environmental noise. Compared with other models, it can detect the category and position of smoke targets more accurately. It has learned more complete and detailed small target features, which greatly alleviates the problem of missed detection.

Table 4. The Visual comparison of different models on AI For Humankind dataset.the a column is the detection result without obvious interference information, the b column is the detection result under strong light illumination, the column c presents detection result under the scene with white clouds in the picture background, and the column d presents detection result under the scene with fog in the picture background.

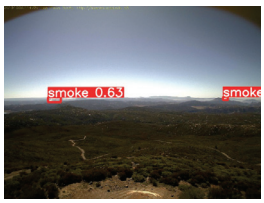
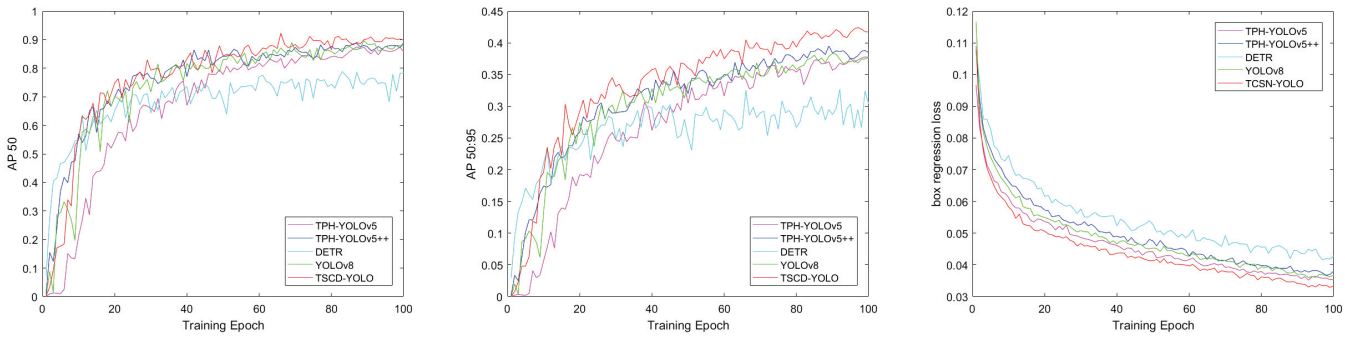
Model	**Detection Results**			
TPH-YOLOv5				
TPH-YOLOv5++				
DETR				
YOLOv8				
TCSN-YOLO				
	(a)	(b)	(c)	(d)

Figure 6 demonstrates the performance comparison with other mainstream methods during training, where Figure 6a–c exhibit the performance curves of AP50, AP50:90, and box regression loss metrics, respectively. To visually observe the performance of the comparison model, we present the training process during the initial 100 epochs. TSCD-YOLO achieves superior performance across all key metrics, demonstrating its enhanced detection capability. Furthermore, its rapid convergence significantly enhances training efficiency while conserving computational resources. The results verify the practicality of the model in actual deployment.



(**a**) (**b**) (**c**)

Figure 6. Comparison with other mainstream methods. (**a**) AP50 curves. (**b**) AP50:90 curves. (**c**) box regression loss curves.

4.6. Generalization Experiment

We verifies the generalization ability of the TSCN-YOLO model on the AI For Humankind and FlgLib datasets [36]. The experimental setup is identical to that employed in the previous experiment. The results are presented in Table 5.

The AI For Humankind dataset included 2194 bounding box annotated images HPWREN Cameras captured on remote mountain tops in Southern California, including images taken under different backgrounds (clouds, strong light, fog, etc.). The smoke images included in this dataset are annotated in the Pascal VOC annotation format. FIgLib was developed and is maintained by Hans-Werner Braun at The University California San Diego (UCSD). The objective is to support neural network researchers in advancing early fire detection systems and fundamental research on fire initiation mechanisms. The dataset contains 24,800 high-resolution images of 1536 × 2048 or 2048 × 3072 pixels, covering a variety of terrains such as mountains, deserts, and coastal regions, and includes temporal sequences spanning an entire year to capture seasonal changes.

Table 5. Generalization performance of the TCSN-YOLO model evaluated on the FIgLib dataset.

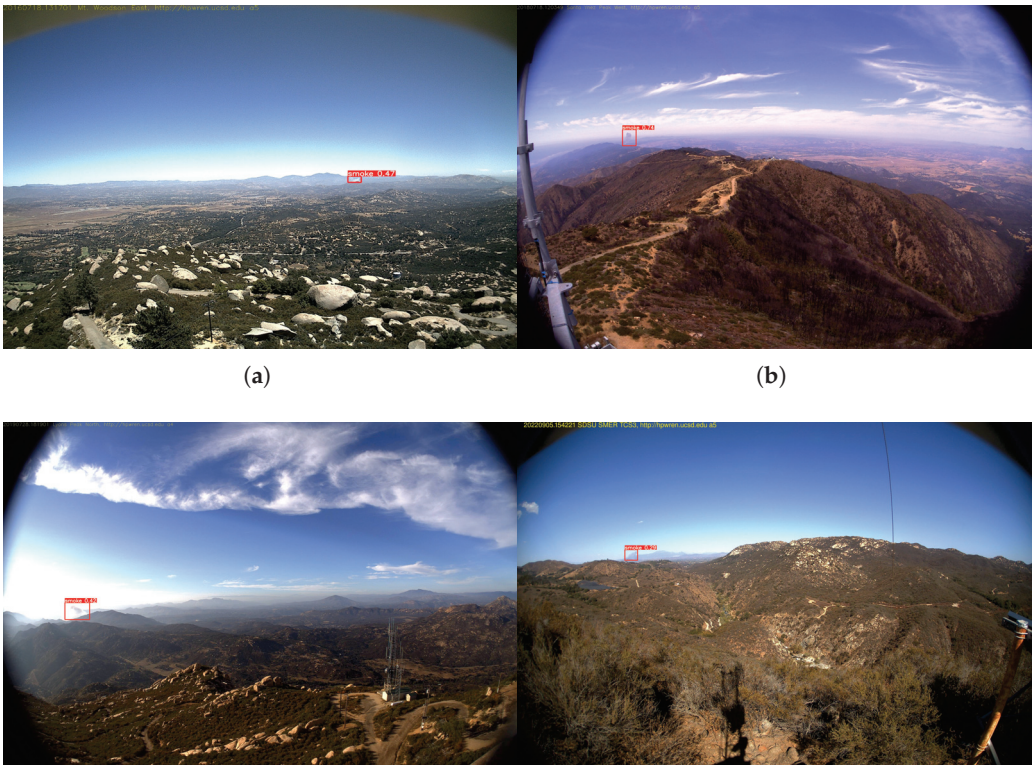
Model	AP_{50} (%)	AP_{75} (%)	AP_S (%)	AP_M (%)
AI For Humankind	90.3	40.4	37.1	61.6
FlgLib	88.0	38.2	35.8	61.1

On the two distinct datasets, the experimental results confirm that TCSN-YOLO maintains high accuracy across both datasets, demonstrating exceptional resilience to interference and robust feature extraction capabilities. The model consistently achieves high AP and AP50:95 values under diverse conditions, indicating strong performance stability for small targets in complex scenes. This validates its suitability for challenging small-target smoke detection applications.

The visualization results on the FlgLib datasets are shown in Figure 7. We selected representative images from the FlgLib dataset to validate the accuracy of our TCSN-YOLO architecture across different fire scenarios. These images contain scenes with a variety of interfering backgrounds. For example, in Image 3, strong light blurs the outline of the smoke, while in Image 4, the color and shape of the smoke closely resemble the background mountains. The visualization results confirm TCSN-YOLO's robust detection competence for small objects across diverse complex scenarios. Its consistent performance on unseen datasets further affirms strong generalization capabilities.

The visualization shows that our proposed model can accurately identify small smoke targets in the image and is robust in scenes with various interference backgrounds (as illustrated in the 3rd and 4th images of Figure 7). Its consistent performance on unseen datasets further affirms its strong generalization capabilities.

Consequently, ablation, comparative, and generalization experiments demonstrate that TCSN-YOLO surpasses mainstream small object detectors while exhibiting robust generalization capabilities. The model sustains high detection precision and stability in challenging scenarios, including complex backgrounds and small targets.



(**a**) (**b**)

(**c**) (**d**)

Figure 7. The visualization results on FlgLib datasets. (**a**) wildfire scenes, stones. (**b**) wildfire scenes, roads, mountains. (**c**) wildfire scenes, strong light. (**d**) wildfire scenes, trees, mountains.

5. Discussion

Aiming at the challenge of detecting small smoke in complex background, we propose TCSN-YOLO, a model based on YOLOv8. To begin, we propose the trident fusion (TF), which receives more shallow features to identify small-sized smoke. Then, to address the difficulty of capturing global dependencies of small smoke, CAM is introduced to optimize the feature extraction capability in different directions in the feature map. Moreover, we substitute the max pooling in the SPPF module with Softpool, which minimizes information loss during the pooling process while preserving the functionality of the pooling layer. Finally, NWD metric be embedded into the loss function of our detector to distinguish

positive and negative samples under the same threshold. We attribute performance gains to each architectural innovation through a series of ablation experiments and conduct a comprehensive comparison with established classical algorithms. The TCSN-YOLO achieves 37.1% APs, 90.3% AP50, and 40.4% AP50:95, surpassing existing methods and demonstrating its effectiveness. The TCSN-YOLO demonstrates excellent anti-interference capabilities and accuracy, making it is very effective for small target fire smoke detection scenarios that require high-precision recognition and robustness.

Although TCSN-YOLO performs effectively in detecting small smoke targets, interference factors such as sunlight exposure at specific angles or low smoke concentration can lead to misjudgments. Future work will focus on studying the characteristic differences between smoke and these interfering factors to enhance the algorithm's robustness. Additionally, while the complex structure of TCSN-YOLO provides powerful capabilities, it introduces challenges in optimization. Increasing the number of parameters imposes greater computational demands on the hardware, potentially limiting the model's applicability for real-time deployment. Furthermore, although we have expanded the dataset, the number is still relatively small. Smaller datasets have fewer smoke features and insufficient scene coverage, which brings the risk of overfitting. Next, we will collect smoke data from more scenarios to improve the generalization ability of the model. Future research should prioritize using cross-modal distillation technology, a multimodal teacher model composed of text and images guides a unimodal student model to learn cross-modal correlation features. This reduces the number of model parameters, enabling the student model to achieve near-multimodal perception capabilities.

Author Contributions: Conceptualization, C.Y.; software, C.Y.; supervision, Z.J.; visualization, W.H. and W.G.; writing—original draft, C.Y.; writing—review and editing, Z.J. and W.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Science Foundation of China grant number 62375129.

Data Availability Statement: The dataset referenced is publicly available (see [35,36]). The experimental data supporting the findings of this study are available from the first author (C.Y.) upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Sathishkumar, V.E.; Cho, J.; Subramanian, M.; Naren, O.S. Forest fire and smoke detection using deep learning-based learning without forgetting. *Fire Ecol.* **2023**, *19*, 9. [CrossRef]
2. Kim, S.Y.; Muminov, A. Forest fire smoke detection based on deep learning approaches and unmanned aerial vehicle images. *Sensors* **2023**, *23*, 5702. [CrossRef] [PubMed]
3. Çınarer, G. Hybrid Backbone-Based Deep Learning Model for Early Detection of Forest Fire Smoke. *Appl. Sci.* **2025**, *15*, 7178. [CrossRef]
4. Kisantal, M.; Wojna, Z.; Murawski, J.; Naruniec, J.; Cho, K. Augmentation for small object detection. *arXiv* **2019**, arXiv:1902.07296.
5. Geetha, S.; Abhishek, C.S.; Akshayanat, C.S. Machine vision based fire detection techniques: A survey. *Fire Technol.* **2021**, *57*, 591–623. [CrossRef]
6. Prema, C.E.; Suresh, S.; Krishnan, M.N.; Leema, N. A novel efficient video smoke detection algorithm using co-occurrence of local binary pattern variants. *Fire Technol.* **2022**, *58*, 3139–3165. [CrossRef]
7. Filonenko, A.; Hernández, D.C.; Jo, K.H. Fast smoke detection for video surveillance using CUDA. *IEEE Trans. Ind. Inform.* **2017**, *14*, 725–733. [CrossRef]
8. Alamgir, N.; Nguyen, K.; Chandran, V.; Boles, W. Combining multi-channel color space with local binary co-occurrence feature descriptors for accurate smoke detection from surveillance videos. *Fire Saf. J.* **2018**, *102*, 1–10. [CrossRef]
9. Li, Z.; Mihaylova, L.S.; Isupova, O.; Rossi, L. Autonomous flame detection in videos with a Dirichlet process Gaussian mixture color model. *IEEE Trans. Ind. Inform.* **2017**, *14*, 1146–1154. [CrossRef]

10. Krstinić, D.; Stipaničev, D.; Jakovčević, T. Histogram-based smoke segmentation in forest fire detection system. *Inf. Technol. Control* **2009**, *38*, 237–244.
11. Sultan, T.; Chowdhury, M.S.; Safran, M.; Mridha, M.F.; Dey, N. Deep learning-based multistage fire detection system and emerging direction. *Fire* **2024**, *7*, 451. [CrossRef]
12. Wang, J.; Zhang, X.; Zhang, C. A lightweight smoke detection network incorporated with the edge cue. *Expert Syst. Appl.* **2024**, *241*, 122583. [CrossRef]
13. Ba, R.; Chen, C.; Yuan, J.; Song, W.; Lo, S. SmokeNet: Satellite smoke scene detection using convolutional neural network with spatial and channel-wise attention. *Remote Sens.* **2019**, *11*, 1702. [CrossRef]
14. Hashemzadeh, M.; Farajzadeh, N.; Heydari, M. Smoke detection in video using convolutional neural networks and efficient spatio-temporal features. *Appl. Soft Comput.* **2022**, *128*, 109496. [CrossRef]
15. Majid, S.; Alenezi, F.; Masood, S.; Ahmad, M.; Gündüz, E.S.; Polat, K. Attention based CNN model for fire detection and localization in real-world images. *Expert Syst. Appl.* **2022**, *189*, 116114. [CrossRef]
16. Khan, S.; Muhammad, K.; Hussain, T.; Del Ser, J.; Cuzzolin, F.; Bhattacharyya, S.; Akhtar, Z.; de Albuquerque, V.H.C. Deepsmoke: Deep learning model for smoke detection and segmentation in outdoor environments. *Expert Syst. Appl.* **2021**, *182*, 115125. [CrossRef]
17. Hosseini, A.; Hashemzadeh, M.; Farajzadeh, N. UFS-Net: A unified flame and smoke detection method for early detection of fire in video surveillance applications using CNNs. *J. Comput. Sci.* **2022**, *61*, 101638. [CrossRef]
18. Zhu, X.; Lyu, S.; Wang, X.; Zhao, Q. TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11 October 2021; pp. 2778–2788.
19. Zhao, Q.; Liu, B.; Lyu, S.; Wang, C.; Zhang, H. Tph-yolov5++: Boosting object detection on drone-captured scenarios with cross-layer asymmetric transformer. *Remote Sens.* **2023**, *15*, 1687. [CrossRef]
20. Amjad, A.; Huroon, A.M.; Chang, H.T.; Tai, L.C. Dynamic fire and smoke detection module with enhanced feature integration and attention mechanisms. *Pattern Anal. Appl.* **2025**, *28*, 81. [CrossRef]
21. Bahhar, C.; Ksibi, A.; Ayadi, M.; Jamjoom, M.M.; Ullah, Z.; Soufiene, B.O.; Sakli, H. Wildfire and smoke detection using staged YOLO model and ensemble CNN. *Electronics* **2023**, *12*, 228. [CrossRef]
22. Kwak, D.K.; Ryu, J.K. A study on the dynamic image-based dark channel prior and smoke detection using deep learning. *J. Electr. Eng. Technol.* **2022**, *17*, 581–589. [CrossRef]
23. Jeong, M.; Park, M.; Nam, J.; Ko, B.C. Light-Weight Student LSTM for Real-Time Wildfire Smoke Detection. *Sensors* **2020**, *20*, 5508. [CrossRef] [PubMed]
24. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 3–19.
25. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-End Object Detection with Transformers. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23 August 2020; pp. 213–229.
26. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 17 March 2021; pp. 10012–10022.
27. Safarov, F.; Muksimova, S.; Kamoliddin, M.; Cho, Y.I. Fire and Smoke Detection in Complex Environments. *Fire* **2024**, *7*, 389. [CrossRef]
28. Rohra, A.; Yin, B.; Bilal, H.; Kumar, A.; Ali, M.; Li, Y. MSFFNet: Multi-scale feature fusion network with semantic optimization for crowd counting. *Pattern Anal. Appl.* **2025**, *28*, 21. [CrossRef]
29. Singh, K.; Parihar, A.S. Bff: Bi-stream feature fusion for object detection in hazy environment. *Signal Image Video Process.* **2024**, *18*, 3097–3107. [CrossRef]
30. Ghiasi, G.; Lin, T.Y.; Le, Q.V. Nas-fpn: Learning scalable feature pyramid architecture for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16 June 2019; pp. 7036–7045.
31. Achinek, D.N.; Shehu, I.S.; Athuman, A.M.; Fu, X. DAF-Net: Dense attention feature pyramid network for multiscale object detection. *Int. J. Multimed. Inf. Retr.* **2024**, *13*, 18. [CrossRef]
32. Zhang, H.; Zhang, S. Focaler-iou: More focused intersection over union loss. *arXiv* **2024**, arXiv:2401.10525.
33. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.
34. Wang, J.; Xu, C.; Yang, W.; Yu, L. A normalized Gaussian Wasserstein distance for tiny object detection. *arXiv* **2021**, arXiv:2110.13389.

35. HPWREN/AI for Mankind. Available online: https://github.com/aiformankind/wildfire-smoke-dataset (accessed on 20 November 2024).
36. Dewangan, A.; Pande, Y.; Braun, H.W.; Vernon, F.; Perez, I.; Altintas, I.; Cottrell, G.W.; Nguyen, M.H. FIgLib & SmokeyNet: Dataset and deep learning model for real-time wildland fire smoke detection. *Remote Sens.* **2022**, *14*, 1007.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

remote sensing

MDPI

Article

LENet: A Semantic Segmentation Network for Complex Landforms in Remote Sensing Imagery via Axial Semantic Modeling and Deformation-Aware Compensation

Yaning Liu [1], Jing Ren [2,*], Jiakun Wang [2], Shaoda Li [1], Rui Chen [2], Dongsheng Zhong [1], Wei Zhao [2], Aiping Yang [2] and Ronghao Yang [1]

[1] Department of Surveying & Mapping Engineering, College of Earth and Planetary Sciences, Chengdu University of Technology, Chengdu 610059, China; liuyaning@stu.cdut.edu.cn (Y.L.); lisd@cdut.edu.cn (S.L.); zdscaptain@stu.cdut.edu.cn (D.Z.); yangronghao@cdut.edu.cn (R.Y.)
[2] The Seventh Geological Brigade of Sichuan Province, Leshan 614000, China; 13508131369@163.com (J.W.); chenrui_abc@163.com (R.C.); 13981364893@139.com (W.Z.); 13981321091@139.com (A.Y.)
* Correspondence: 15882126124@163.com

Highlights

What are the main findings?

- Proposes LENet, a novel semantic segmentation network for complex landforms.
- Achieves state-of-the-art performance on PKLD and GVLM datasets with high robustness.

What are the implications of the main findings?

- Introduces axial semantic modeling to enhance long-range contextual dependencies.
- Designs a Feature Expert Compensator (FEC) for deformation-aware intra-class fusion.
- Employs Cross Sparse Attention (CSA) to suppress background noise and refine details.

Abstract

Accurate semantic segmentation of complex landforms in remote sensing imagery is hindered by pronounced intra-class heterogeneity, blurred boundaries, and irregular geomorphic structures. To overcome these challenges, this study presents LENet (Landforms Expert Segmentation Net), a novel segmentation network that combines axial semantic modeling with deformation-aware compensation. LENet follows an encoder–decoder framework, where the decoder integrates three key modules: the Expert Enhancement Block (EEBlock) for capturing long-range dependencies along axial directions; the Feature Expert Compensator (FEC) employing deformable convolutions with channel–spatial decoupled weights to emphasize ambiguous intra-class regions; and the Cross-Sparse Attention (CSA) mechanism that suppresses background noise via multi-rate sparsity masks and enhances intra-class consistency through cosine-similarity weighting. Experiments conducted on the PKLD plateau karst and GVLM landslide datasets demonstrate that LENet achieves IoU scores of 70.39% and 80.95% and Recall values of 83.33% and 91.38%, surpassing eight state-of-the-art methods. These results confirm that LENet effectively balances global contextual understanding and local detail refinement, providing a robust and accurate solution for complex landform segmentation in remote sensing imagery.

Keywords: complex landforms; semantic segmentation; axial semantic modeling; deformation-aware compensation; intra-class information compensation

1. Introduction

Remote sensing imagery plays a critical role in large-scale terrain analysis, natural disaster monitoring, and environmental assessment [1]. As an essential task in the intelligent interpretation of remote sensing imagery, semantic segmentation aims to assign explicit semantic category labels to each pixel, thereby enabling accurate identification of surface objects and landform types as well as modeling their spatial distribution [2]. This technique has been widely applied to land use classification [3], crop yield estimation [4], urban expansion monitoring [5], and ecological environment assessment [1], becoming one of the core approaches driving intelligent geographic information analysis.

Among various application scenarios, the semantic segmentation of complex landforms (e.g., karst terrains, landslide areas, and post-earthquake regions) is increasingly becoming a research hotspot. In China, for instance, the exposed or near-surface soluble rock covers an area of more than 1.3 million km^2, and the widely distributed karst landforms play a vital role in geological surveys, water resource protection, and infrastructure site selection [6,7]. At the same time, landslides, as one of the most common and destructive natural hazards worldwide, also exhibit typical characteristics of complex landforms. Their irregular boundaries and intricate formation mechanisms impose greater challenges on semantic segmentation models [8,9]. Therefore, developing a high-precision semantic segmentation method tailored for complex landforms holds significant scientific value and promising application potential for advancing remote sensing-based geoscientific interpretation.

Compared with natural scenes such as forests and farmlands [1,4], or urban scenes such as roads [10] and buildings [3,5], complex landform remote sensing images face significant challenges in semantic segmentation tasks. First, complex landforms exhibit substantial intra-class variations in elevation, scale, spectral reflectance, and texture, which weakens the model's ability to achieve semantic consistency within the same category. Second, some landform classes share high similarity with the background in terms of spectral characteristics or spatial distribution, making it difficult for traditional feature extraction methods to effectively distinguish landform categories from the background. Finally, complex landform regions often lack clear boundaries, present irregular shapes with vast spatial extents, and may even evolve over time, further aggravating boundary recognition errors and inaccurate regional delineation (as illustrated in Figure 1).

Traditional remote sensing semantic segmentation methods, such as threshold-based approaches [11] and machine learning techniques [12], often rely on fixed parameter settings. Consequently, their performance is usually unsatisfactory when dealing with regions under varying illumination conditions and complex textures. In recent years, deep learning has achieved remarkable progress in remote sensing semantic segmentation, with Convolutional Neural Networks (CNNs) being widely applied in environmental remote sensing [13], geological remote sensing [14], and urban remote sensing [3,5]. For example, in tasks such as urban building and road extraction [3,5,10,15], small object detection in remote sensing imagery [16], and urban change detection [17], CNN-based models leverage multi-scale receptive fields and spatial pyramid structures to effectively capture spatial context, delivering high segmentation accuracy in structured scenes. However, their receptive fields are inherently constrained by local convolutional kernels, making it difficult to model long-range spatial dependencies. Moreover, CNNs are insufficiently sensitive to features with significant scale variations and blurred boundaries, which can lead to intra-class feature loss and background misclassification in boundary regions. To overcome these limitations, Transformer architectures have been introduced into remote sensing image segmentation owing to their superior global modeling capability [18]. Representative methods, such as Swin Transformer [19], Mix Vision Transformer [20], and DeBifomer [21], employ self-

https://doi.org/10.3390/rs18010059

attention mechanisms to enable cross-region information interaction and exhibit strong potential in multi-scale feature aggregation, achieving excellent performance when directly applied to remote sensing semantic segmentation. Nevertheless, for complex landform remote sensing imagery, although attention mechanisms can effectively capture global relationships among target regions, they still fail to address the critical issue of local detail feature loss.

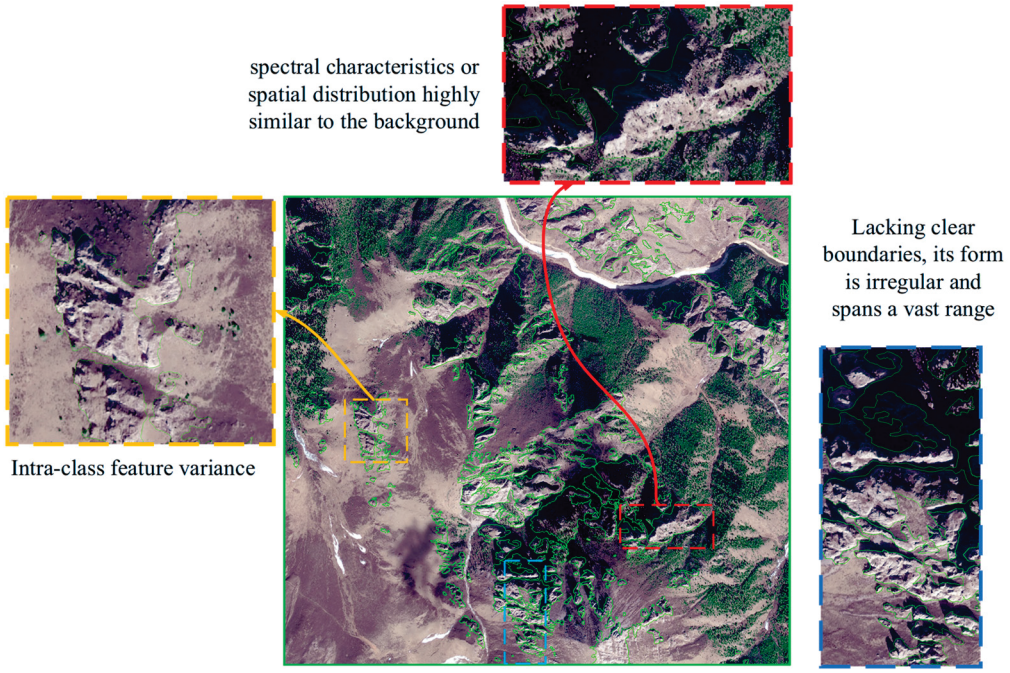


Figure 1. Examples of challenging scenarios in remote sensing image semantic segmentation (selected from the karst dataset PKLD and the public landslide dataset GVLM).

To leverage the complementary advantages of CNNs and Transformers, several studies have explored hybrid architectures that integrate the two. Representative approaches include DC-Swin [22] and ST-Unet [23], which preserve the Transformer's strength in modeling long-range dependencies while retaining the CNN's ability to capture local details. However, these methods still pay insufficient attention to indistinguishable intra-class details, and when background and target features exhibit high similarity, they may lead to partial loss of intra-class feature information.

In order to improve the segmentation accuracy of complex terrain remote sensing images while maintaining the integrity of intra-class features, this paper proposes a novel semantic segmentation network, LENet (landforms Expert segmentation Net), based on axial semantic modeling and deformation-aware compensation. Building on previous work integrating CNNS and Transformers. Inspired by the cognitive process of human experts in recognizing complex geomorphic regions, LENet mimics the way experts utilize domain knowledge: they first combine global and local feature distributions to filter most of the background interference, then retain regions rich in geomorphology features, and finally pay special attention to indistinguishable fuzzy intra-class regions. Therefore, in complex landform image segmentation, the key challenge is to quickly eliminate background interference while focusing on complex, fuzzy regions in complex landform classes, which is crucial to improve segmentation performance. Among them, the visualization of the LENet imitating the workflow of human experts and the working principle of each component is shown in Figure 2.

https://doi.org/10.3390/rs18010059

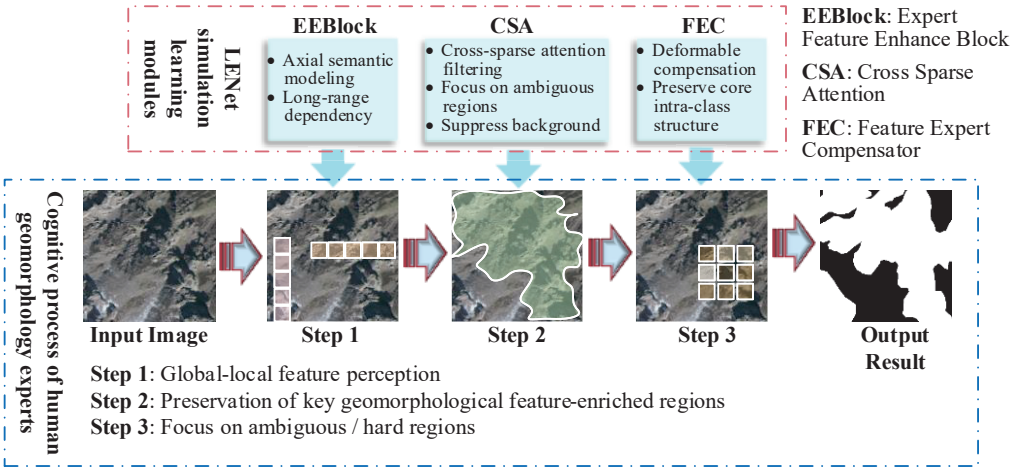


Figure 2. LENet performs visualizations that simulate the process of human experts identifying features and how each component works, where colored or rectangular areas represent demonstrations of the scope of action.

LENet is built upon a state-of-the-art (SOTA) encoder–decoder architecture, with its key innovation lying in the design of the decoder. As the crucial stage for feature reconstruction, the decoder determines how to effectively restore category information from feature maps rich in semantic representations, which remains a major challenge in complex landform segmentation tasks. To address the wide spatial extent of complex geomorphological regions, we design an expert-enhanced axial semantic modeling module in the decoder to capture both horizontal and vertical contextual information, thereby adapting to large-scale spatial feature distributions. To mitigate noise interference caused by the similarity between background and target features, a cross-sparse attention mechanism is introduced to filter out redundant background information. Furthermore, to handle the rich intra-class scale variations in complex landform categories, a feature expert compensator is proposed to emphasize critical intra-class regions, thereby enhancing semantic consistency within categories. The main contributions of this work can be summarized as follows:

(1) We propose a novel remote sensing semantic segmentation network, LENet, which integrates axial feature modeling with expert-inspired intra-class information identification. By combining an attention-based encoder with strong global modeling capability and a hybrid decoder capable of capturing subtle local features and filtering key information, LENet addresses the challenge of insufficient intra-class recognition integrity in complex landform segmentation.
(2) We design an expert-inspired multi-scale feature learning decoder that mimics the process of expert judgment in distinguishing complex geomorphological features. The decoder incorporates an Expert Feature Enhancement Block (EEBlock), a Feature Expert Compensator (FEC), and a Cross Sparse Attention (CSA) module to improve category integrity and intra-class feature accuracy.
(3) We validate the effectiveness of LENet on a landslide dataset and a karst dataset. Experimental results demonstrate that LENet achieves superior performance across multiple evaluation metrics, confirming its robustness and advancement in complex landform segmentation tasks.

2. Related Works

2.1. Remote Sensing Image Semantic Segmentation

Semantic segmentation can be regarded as a pixel-level extension of image classification, with the goal of assigning a semantic category label to each pixel in the image [3].

https://doi.org/10.3390/rs18010059

Traditional semantic segmentation methods typically rely on manually annotated pixel-level labels and employ deep neural networks to learn the mapping between pixel features and semantic categories. CNN-based approaches have achieved remarkable progress in remote sensing image analysis [22–25], by constructing multi-layer convolutional architectures to extract and classify local features on a per-pixel basis. With further research, many studies have attempted to enhance contextual modeling capability and enlarge the receptive field to improve semantic understanding among pixels. U-Net introduces skip connections to integrate multi-scale contextual information [24], becoming a classical model for semantic feature enhancement through multi-scale fusion. SegNeXt employs large-kernel strip convolutions to expand the receptive field [26]. BiSeNet proposes a dual-branch structure to balance spatial detail preservation and semantic understanding efficiency [27], while EncNet enhances semantic category-related feature representation through a context encoding module [28].

Since the introduction of Transformer architectures into visual tasks, their reliance on self-attention mechanisms has demonstrated remarkable advantages in modeling long-range dependencies [18,29,30], gradually becoming a pivotal direction in the design of semantic segmentation models. However, Transformers incur substantial computational costs when processing high-resolution images, which limits their practical deployment in remote sensing applications. To address this issue, the researchers proposed a series of lightweight optimization strategies: In addition to the classic moving window mechanism adopted by Swin Transformer [19], Biformer utilizes a novel dynamic sparse attention mechanism achieved through a two-layer routing approach, enabling more flexible and content-aware computational allocation [31]; RMT introduces a spatial decay matrix and designs a spatially constrained attention decomposition structure to enhance long-range modeling efficiency [32]; Zhu et al. integrate CNNs with Transformers to extract high-frequency details and low-frequency semantic features separately, achieving a unified representation of global context and fine-grained structures [33].

Although the aforementioned methods achieve excellent performance in natural image semantic segmentation, they often struggle when directly applied to remote sensing imagery due to the unique challenges of such data, including high intra-class variance, significant scale variations, blurred object boundaries, and complex backgrounds. In addition, remote sensing images typically possess ultra-high resolution, which further increases model parameters and computational costs, thereby limiting the practicality and scalability of these approaches.

To address the unique challenges of remote sensing imagery, numerous scholars have proposed targeted design strategies. For instance, in urban remote sensing scenarios, common issues include high inter-class similarity, large intra-class variation, and dense small objects [5,22,34]. Wang et al. introduced a parallel window attention structure to enhance the spatial modeling capability of Swin Transformer [35], thereby improving urban remote sensing image segmentation performance. Rau et al. incorporated DEM auxiliary information to achieve hierarchical segmentation for multimodal landslide disasters [36]. Cui et al. developed a cross-modal transfer learning framework for semantic segmentation [37], introducing channel and spatial discretization losses to alleviate inter-modal feature conflicts and redundancy, thus enhancing modality-cooperative modeling. He et al. employed prompt learning mechanisms and explicit semantic grouping strategies to adapt frozen pre-trained models to multimodal downstream tasks [38]. Ma et al. combined the strengths of CNNs and Transformers to propose a hybrid architecture that improves multimodal fusion efficiency [15]. In addition, Zhang et al. utilized infrared remote sensing imagery for vehicle detection [39], Zheng et al. designed a foreground-aware relational modeling network to tackle scale variations and foreground class imbalance in remote

https://doi.org/10.3390/rs18010059

sensing images [40], and Chen et al. introduced a prompt-driven mechanism based on the SAM architecture for instance segmentation in remote sensing imagery [41].

To enhance the spatial sampling capability of convolutional neural networks, Dai et al. proposed Deformable Convolutional Networks (DCN) [42], which learn unsupervised offsets from target tasks to expand the spatial sampling positions of convolutional layers, demonstrating strong potential in remote sensing semantic segmentation. Yu et al. leveraged deformable convolutions to achieve global and deformation-aware feature extraction [10], improving road extraction performance in remote sensing imagery. Dong et al. introduced a multi-scale deformable attention module [43], utilizing a larger deformable receptive field to adapt to remote sensing targets of diverse shapes and sizes, thereby generating more precise attention maps. Hu et al. proposed a multi-scale deformable self-attention mechanism to reduce the partial loss of linearly intertwined road information in remote sensing images [44], effectively enhancing the saliency of road features relative to the surrounding environment. Although deformable convolutions have been widely applied in remote sensing semantic segmentation, most applications focus on complex object shapes or objects densely intertwined with the background, while their substantial potential for handling intra-class feature variations within target categories remains underexplored.

In summary, although recent advances in CNNs and Transformer architectures have significantly promoted the development of remote sensing semantic segmentation, most existing studies focus on man-made or natural land cover types with regular boundaries and clear structures, such as urban buildings, roads, and agricultural areas. Effective modeling strategies remain lacking for complex landforms, which exhibit irregular textures, atypical features, blurred boundaries, and high intra-class variance. While deformable convolutions have demonstrated remarkable capabilities in road recognition tasks, their potential for learning critical intra-class regions has yet to be fully explored. Consequently, directly applying existing methods to complex landform remote sensing imagery often fails to achieve satisfactory segmentation performance.

2.2. Semantic Segmentation of Complex Landforms

Semantic segmentation of complex landform regions in remote sensing imagery is a highly challenging task, due to the spatial heterogeneity of multiple landform types, blurred boundaries, significant scale variations, and high intra-class variance. Areas exhibiting typical complex landform characteristics, such as landslide-prone zones, plateau karst regions, and desert hilly landscapes, are of particular interest, as accurate semantic understanding and fine-grained segmentation in these regions play a crucial role in disaster monitoring, resource surveying, and ecological assessment.

In geohazard-related complex landforms, remote sensing monitoring and semantic segmentation of landslide-prone areas have become a major research focus. Landslide occurrences are typically accompanied by abrupt changes in surface morphology, posing significant threats to human life and property, making their mapping and dynamic change detection highly practical. Zhang et al. proposed a prototype-guided region-aware progressive learning method based on multi-scale target domain adaptation to address cross-domain modeling of landslide regions in large-scale remote sensing imagery [8]. Lu et al. designed a lightweight network, MS2LandsNet, specifically for landslide landform features [14]; by reducing the number of channels and optimizing the network architecture, combined with multi-scale feature fusion and channel attention mechanisms, the model's performance in landslide detection tasks was significantly improved. Şener et al. developed LandslideSegNet [45], which integrates an encoder–decoder residual structure with spatial–channel attention mechanisms to enhance early identification of potential landslide regions. Soares et al. evaluated the generalization performance of automated

https://doi.org/10.3390/rs18010059

landslide mapping models across three landform types incorporating NDVI and DEM data, demonstrating that NDVI information helps mitigate overfitting and improves prediction balance [46].

Research on remote sensing segmentation of non-hazard-related complex landforms has also been continuously expanding. Huang et al. addressed the challenge of multi-scale representation of landscape landform features by proposing the SwinClustering framework based on Swin Transformer [19], enhancing the model's ability to jointly capture spatial structure and semantic information [47]. Cheng et al. designed an improved semantic segmentation framework that employs statistical preprocessing to remove invalid image patches during training [48], thereby improving the model's capacity to capture local details of fracture-like features. Yu et al. introduced SegRoadv2 for road segmentation [10], incorporating deformable self-attention and grouped deformable convolution structures to enhance the perception of irregular road boundaries. Goswami et al. utilized a U-Net architecture to construct segmentation models for typical landform types, including deserts, forests, and mountainous regions [49]. Zhou et al. proposed a network based on a dense attention residual pyramid fusion structure for Lithological Unit Classification (LUC) [50], a task of significant relevance in geological resource exploration. While these methods have achieved remarkable results across various complex landform types, they differ from the focus of this study, which aims to develop a generalized segmentation model capable of handling diverse complex landforms. Specifically, this work proposes a semantic segmentation method for complex landform remote sensing imagery that not only improves segmentation accuracy but also preserves intra-class feature integrity.

3. Methodology

3.1. Framework Overview

The overall architecture of LENet (as illustrated in Figure 3) adopts the widely used encoder–decoder structure for semantic segmentation tasks [26,50,51]. For the input raw imagery, the encoder progressively downsamples the feature maps to effectively capture multi-scale contextual information, producing abstract semantic features at different scales S_i (i = 1, 2, 3, 4). The decoder then progressively upsamples and fuses these features, combining high-level semantic information from the encoding stage with low-level spatial details, thereby gradually restoring the spatial structure of the target. This process achieves a balance between global semantic modeling and local detail preservation. Such characteristics make the encoder–decoder framework particularly suitable for complex landform segmentation, as it maintains high-level semantic abstraction while retaining essential local details.

To accommodate the need for modeling long-range dependencies in complex landforms, the encoder of LENet adopts Swin Transformer as its backbone. Compared with traditional convolutional neural networks [19], Swin Transformer offers superior global dependency modeling capabilities. Moreover, its shifted window attention mechanism significantly reduces computational and memory costs, achieving a balance between efficient processing and effective global context modeling.

The decoder of LENet consists of two core components: the EEBlock and the Decoder Block (as illustrated in Figure 4). The EEBlock employs axial semantic modeling to extract semantic features along both horizontal and vertical directions, enabling efficient long-range dependency modeling and enhancing the representation of morphologically irregular and spatially extensive landform categories. The Decoder Block comprises two parallel modules: the FEC and CSA. FEC reinforces intra-class feature representation through cross-scale feature fusion and adaptive compensation, mitigating information loss and inconsistencies caused by significant intra-class variations. CSA models the correlations

https://doi.org/10.3390/rs18010059

between multi-scale features based on cross-sparse attention, enhancing the discriminability of key regions while suppressing background interference. Finally, a cosine similarity mechanism integrates the compensated intra-class semantic information into the selected features, achieving synergistic intra-class compensation and key region enhancement. This design maintains global modeling capabilities while effectively strengthening local detail representation, thereby improving the semantic segmentation accuracy of complex landform regions.

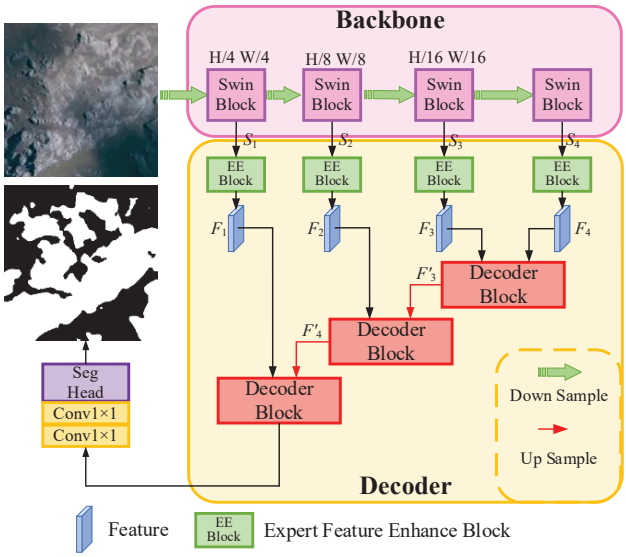


Figure 3. The overall flowchart of LENet. LENet adopts an encoder–decoder structure, and the encoder uses Swin Transformer-tiny by default. The decoder part consists of an Expert Feature Enhancement Block (EEBlock) and a Decoder Block. EEBlock is responsible for performing axial semantic enhancement on the features extracted by different-stage encoders, and then inputting them into the Decoder Block to further restore the spatial structure.

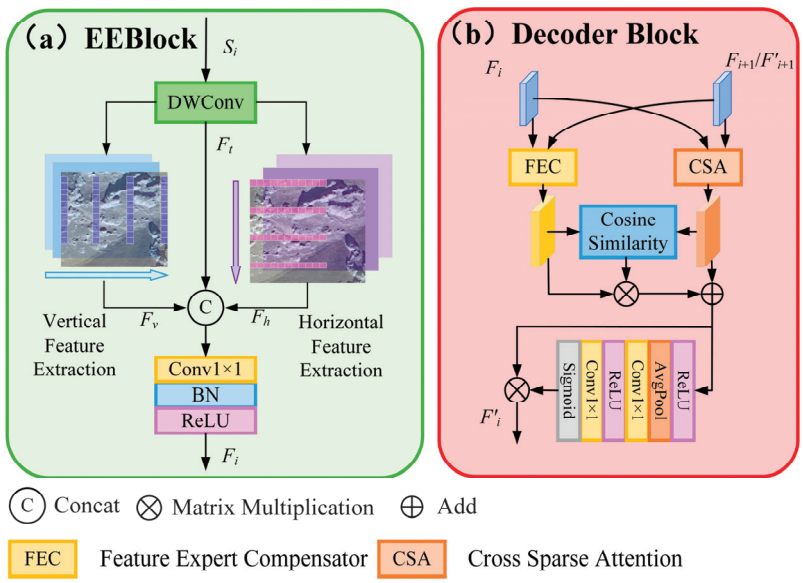


Figure 4. The structural details of EEBlock and Decoder Block. (**a**) The part here represents the structural details of EEBlock, which includes Depthwise Convolution (DWConv) and horizontal and vertical feature extractors. Its structure is designed to enhance the network's ability to model long distances. (**b**) Section presents the structural details of the Decoder Block, which consists of a FEC, CSA, and Cosine Similarity Weighted Part. This structure is used to maintain global modeling while enhancing local important details.

https://doi.org/10.3390/rs18010059

3.2. Expert Feature Enhancement Block (EEBlock)

To address the challenges of extensive spatial coverage and irregular category shapes in complex landform remote sensing imagery, we designed an EEBlock to improve the network's ability to model long-range semantic information, thereby strengthening intra-class long-distance dependencies.

Traditional convolutional networks excel at capturing local features but struggle to model long-range semantic relationships, whereas self-attention mechanisms can achieve long-range modeling but incur significant computational costs as feature map size increases. To leverage the complementary advantages of convolutional networks and self-attention, inspired by MSCAN [26], we designed horizontal and vertical convolutions to simulate self-attention for capturing long-range dependencies along both directions. As illustrated in Figure 4a, the input feature S_i is first processed by a Depthwise Convolution (DWConv) to aggregate local information, producing F_t. The DWConv can be expressed by Equation (1):

$$F_t = \mathrm{DWConv}(S_i) = \sum_{m=0}^{k_h-1} \sum_{n=0}^{k_w-1} S_i \left(h + m - \left[\frac{k_h}{2}\right],\ w + n - \left[\frac{k_w}{2}\right]\right) \cdot W_{c,m,n}^{(k)} \quad (1)$$

where the input feature tensor S_i is a four-dimensional tensor (b, c, h, w). In S_i, $h + m - [k_h/2]$ and $w + n - [k_w/2]$ denote the actual coordinates in the height and width dimensions of the input feature map when the convolution kernel is centered at (h, w). The convolution kernel parameter $W_{c,m,n}^{(k)}$ represents the weight value at position (m, n) of the kernel with size $[k_h, k_w]$ in the c-th channel.

Then, F_t will be processed by a series of parallel vertical and horizontal feature extractors, which, respectively, construct long-range dependency models in the vertical and horizontal directions. Each feature extractor is composed of rectangular DWConvs, which can effectively replace large kernel convolutions using the independent channel convolutional kernel structure of DWConvs, with kernel sizes of [1, 11] and [11, 1], respectively. These parameters were determined through testing, aiming to balance the effective receptive field and computational resource consumption. At the same time, this strip-shaped convolutional kernel design can effectively extract long-structured information in both horizontal and vertical directions. If it were just a regular convolution, multiple layers of stacking or the use of an extremely large kernel would be required to achieve a similar receptive field. After the processing by the feature extractors, F_t will generate horizontal and vertical enhanced features F_v and F_h. By connecting F_t, F_v, and F_h along the channel dimension and applying a 1×1 convolution, the comprehensive feature F_i (i = 1, 2, 3, 4) can be obtained (it enhances the long-range semantic relationships of complex terrain categories in both directions and alleviates the loss of feature correlation due to significant scale changes). The extraction of axial semantic information in the horizontal and vertical directions can refer to Formulas (2) and (3):

$$F_h = \mathrm{DWConv}_{1\times 11}(F_t) \quad (2)$$

$$F_v = \mathrm{DWConv}_{11\times 1}(F_t) \quad (3)$$

3.3. Feature Expert Compensator (FEC)

In complex landform remote sensing imagery, intra-class information loss and inconsistency—such as in karst regions where surface spectral features exhibit minimal contrast with the background and texture boundaries are ambiguous—pose challenges for models to effectively capture deep semantic correlations among pixels of the same class. Traditional convolutional operations often struggle to focus on fine-grained regions with low intra-class discriminability, leading to insufficient feature representation and loss of spatial details. Furthermore, in remote sensing images, multi-scale landform features can

https://doi.org/10.3390/rs18010059

differ significantly in their representation, with the same class exhibiting distinct geometric shapes and texture patterns at different scales. Direct feature fusion in such cases can easily induce semantic misalignment or redundant amplification, further exacerbating semantic information loss across scales. To address these issues, we propose the FEC module, which employs deformable convolutions to emulate the expert identification of hard-to-recognize intra-class regions, thereby enabling more effective cross-scale feature integration and enhancing the model's ability to capture intra-class structural information in complex landforms.

As illustrated in Figure 5, the FEC employs a parallel mechanism that decouples channel and spatial weights, guiding the network to focus on feature channels with higher semantic discriminability while compensating for spatial information in regions that are difficult to learn. First, the input multi-scale features Feature1 and Feature2 are summed to obtain the fused feature F_F. In the channel branch, F_F undergoes adaptive average pooling and adaptive max pooling, followed by concatenation and weighting to generate the channel weight CA. In the spatial branch, F_F is processed by computing both the maximum and average values, which are then concatenated to obtain F_S. F_S is further passed through an additional convolution to produce the spatial offset O_{ff}, and the concatenated feature together with the spatial offset is input into a Deformable convolution (DConv) to perform weighted learning of intra-class spatial information, resulting in the intra-class spatial feature compensation weight F_A. Subsequently, the obtained F_A and CA are multiplied by the original inputs Feature1 and Feature2, respectively, and then summed to integrate semantic information. Adding the residual of F_F yields the Output, representing the fully compensated intra-class feature. Through this mechanism, the FEC balances the channel relationships after multi-scale feature fusion and mitigates intra-class information loss along the spatial dimension, thereby achieving more accurate recognition of intra-class features.

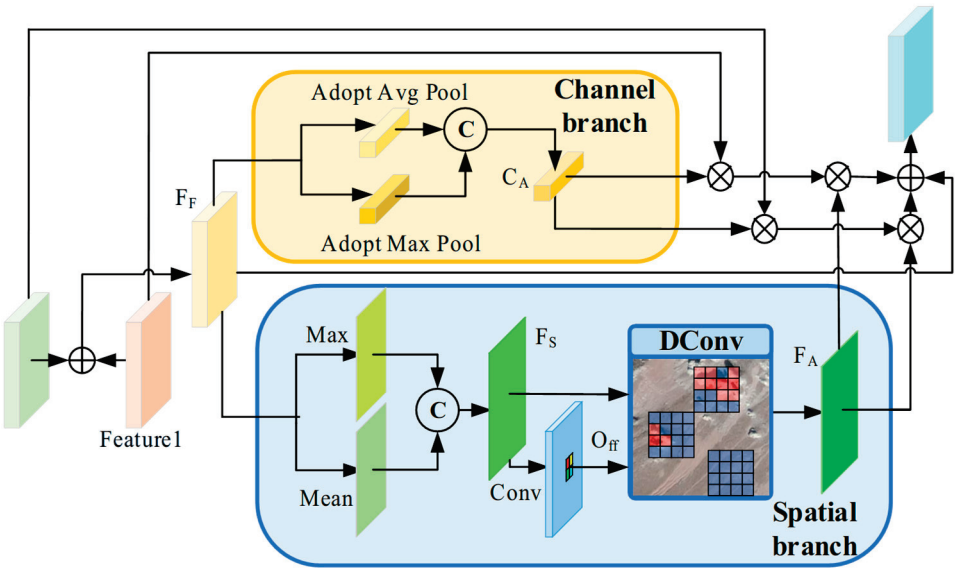


Figure 5. Workflow diagram of the FEC. FEC learns the intra-class difficult-to-classify pixel regions by learning channel weights, spatial information and using deformable convolution, thereby reducing intra-class inconsistency and intra-class information loss.

The generation of the spatial offset and the application of deformable convolution are expressed in Equations (4)~(6):

$$F_s = \text{Concat}(\text{Mean}(F_F), \quad \text{Max}(F_F)) \tag{4}$$

https://doi.org/10.3390/rs18010059

$$O_{ff} = \mathrm{Conv}(F_S) \tag{5}$$

$$F_A = \mathrm{DConv}\left(F_s, O_{ff}\right) \tag{6}$$

The deformable convolution used to generate the spatial compensation weights differs from traditional convolution in that it learns spatial offsets from the spatial feature information through an additional convolutional layer. The input channel number of this convolutional layer is 2 × K (where K is the kernel size, e.g., K = 9 for a 3 × 3 convolution), and each offset at a given position contains both horizontal and vertical components. These offsets are optimized via backpropagation during training, allowing the convolutional kernel to adaptively adjust its sampling locations based on the input features. For a traditional convolution, the output feature map y at position p_0 given an input feature map x is computed as in Equation (7):

$$y(p_0) = \sum_{p_n \in R} w(p_n) \cdot x(p_0 + p_n) \tag{7}$$

here, w represents the convolution kernel weights, and x denotes the weighted sum of the input features within a local neighborhood (p_n). This allows the convolution to focus on details within the neighborhood but does not permit weighted sampling of different regions within it. Deformable convolution addresses this limitation by introducing Δp_n into the convolution, making the sampling locations $p_0 + pn + \Delta p_n$. This enables the kernel to adaptively focus on specific regions within the neighborhood, as shown in Equation (8):

$$y(p_0) = \sum_{p_n \in R} w(p_n) \cdot x(p_0 + p_n + \Delta p_n) \tag{8}$$

3.4. Cross-Sparse Attention (CSA)

In complex landform remote sensing images, landform regions are often accompanied by numerous background features that are similar to the target class. Most feature fusion methods tend to capture these redundant features during global feature modeling, which introduces noise and weakens the representation of key features. To address this issue, we designed a CSA mechanism to reduce the interference from background noise. Traditional cross-attention computes using the full attention weights and cannot filter out irrelevant background noise. Sparse attention, often used to reduce computational complexity in attention mechanisms [52,53], leverages sparse attention maps and has demonstrated effectiveness in tasks such as image denoising [54,55], helping models remove irrelevant noise. Inspired by Top-K sparse attention [56], we propose CSA to filter out irrelevant noise among multi-scale features processed by the expert feature enhancer, as shown in Figure 6.

Unlike traditional cross-attention, which relies on the full attention mask for computation, CSA introduces multiple sparse-rate masks and applies weighted aggregation. Masks with higher weights, corresponding to relevant class information, are retained, while irrelevant background noise is assigned to masks with lower weights. This approach reduces computational complexity and suppresses interference from unrelated features. Such a design effectively diminishes noise in the attention weights, enhances the model's selective learning of semantic information, and thereby improves the discriminative capability for complex landform regions.

Specifically, Feature1 and Feature2 represent the input multi-scale features. Each is first processed through normalization and a standard convolution module (Conv-BN-ReLU) for channel integration. Then, a 1 × 1 convolution layer maps the features to generate the Query (Q), Key (K), and Value (V), which serve as the inputs to CSA. Feature1 generates Q, while Feature2 generates K and V. The attention map is obtained by multiplying Q with K, followed by the addition of relative position encoding to establish the spatial

https://doi.org/10.3390/rs18010059

relationships between features. Subsequently, four sparse-rate mask matrices are applied to selectively filter elements in the attention map, which are then multiplied with V. Finally, the attention maps from different sparse rates are fused via weighted multiplication to produce the output features, which are again processed through normalization and a standard convolution module for non-linear enhancement, thereby improving the representation of relevant features. The CSA can be formally defined as in Equation (9):

$$CSA = \sum_i^4 \omega_i(\text{softmax}(S_i(\frac{QK^T}{\sqrt{d}} + B)))V \quad (9)$$

here, Q represents the query token generated from Feature 1, while K and V represent the key and value tokens generated from Feature2. B denotes the relative position encoding. Si corresponds to the *i*-th (i = 1, 2, 3, 4) sparse-rate mask matrix, and ω_i represents the attention weight associated with the *i*-th sparse-rate mask.

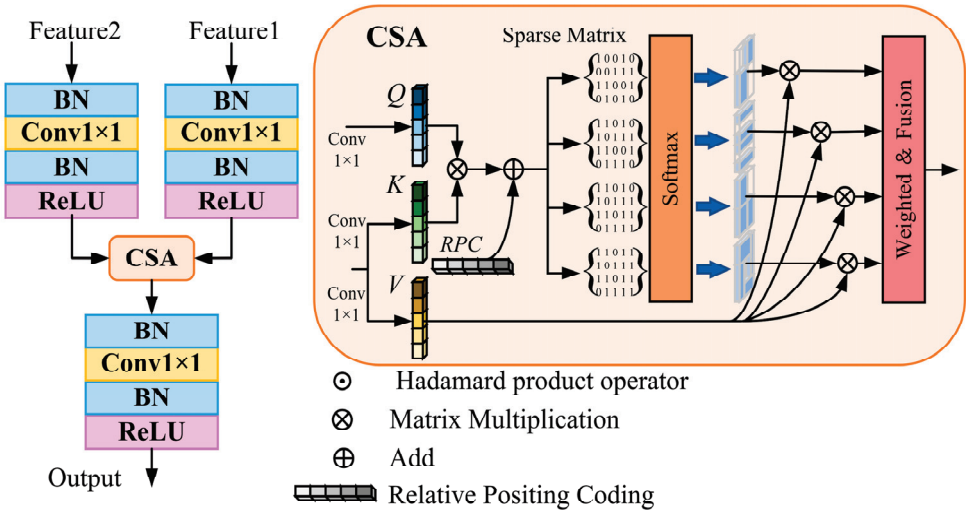


Figure 6. Flowchart of the CSA. On the left side, it represents the complete process of this module, while on the right side, it shows the specific calculation process of cross-sparse attention. CSA can reduce the interference of useless background information by introducing sparse matrices, and normalization and standard convolution groups help this module enhance the feature expression ability.

4. Experiments and Results

4.1. Datasets

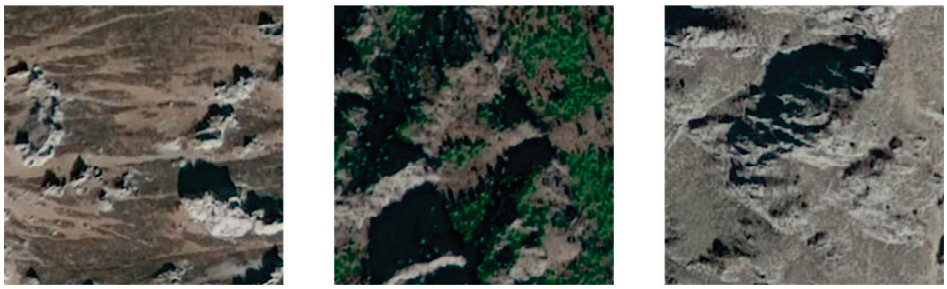
We evaluated LENet on a karst dataset and a landslide dataset. Karst landforms are often associated with faults [57], folds, and other geological structures, exhibiting markedly discontinuous and irregular distributions characteristic of complex landforms. Landslides typically occur in geologically fragile or steep terrain areas, with distributions that are patchy, fragmented, and uneven. Multiple-phase or repeated composite landslides may occur within the same region, resulting in dynamically evolving landform patterns that also exhibit complex landform characteristics.

4.1.1. Karst Landform

Exposed karst landform segmentation in remote sensing imagery is a highly challenging task due to its complex characteristics, including large intra-class variance, uncertain boundary information, and a complex surrounding background, making it an excellent subject for studying complex landform segmentation. We extended the PKLD karst dataset originally created by Zhong et al. [57]. The expanded PKLD contains 17,720 karst images of size 256 × 256 pixels, with a training, validation, and test split of 6:2:2, i.e., 10,632 images for training, and 3544 images each for validation and testing. This area is located in Gongjue

https://doi.org/10.3390/rs18010059

County, Chamdo, Tibet Autonomous Region, China, where the surface karst landforms are mainly composed of limestone peaks and limestone debris accumulations. To account for the wide distribution of karst landforms, the dataset includes images from diverse mountain backgrounds, as shown in Figure 7, including karst on bare mountains (Figure 7a), karst partially covered by vegetation (Figure 7b), and karst in pronounced stone forest regions (Figure 7c).



(**a**) (**b**) (**c**)

Figure 7. Schematic diagrams of remote sensing images of karst landforms under different background types. (**a**) Karst on bare mountains (**b**) Karst partially covered by vegetation (**c**) Karst in pronounced stone forest regions.

4.1.2. Landslide Landform

To evaluate the generalization capability of LENet on complex landform segmentation, we selected landslide landforms as an auxiliary test. Landslide areas are widely distributed and also exhibit the challenging characteristics typical of complex landform segmentation. We employed the publicly available GVLM landslide dataset [8], which contains numerous landslide sites across six continents, including Asia, Africa, North America, South America, Europe, and Oceania, covering diverse countries and regions. This dataset is well-suited for segmentation tasks of landslide-related complex landforms under varying background conditions. For our experiments, we curated 2690 images of size 256 × 256 pixels from GVLM and randomly split them into training, validation, and test sets in a 6:2:2 ratio using a fixed random seed, thereby demonstrating LENet's generalization capability for landslide complex landform segmentation.

4.2. Evaluation Metrics

In this study, we employed four commonly used metrics—Intersection over Union (IoU), Precision, Recall, and F1 score—as the core evaluation criteria to assess the performance of the proposed method.

IoU primarily measures the degree of overlap between the predicted region and the ground truth, serving as an important indicator of overall segmentation accuracy. Its calculation is expressed as follows (Equation (10)):

$$IoU = \frac{TP}{TP + FN + FP} \tag{10}$$

here, True Positives (TP), False Positives (FP), and False Negatives (FN) represent the number of correctly classified pixels, incorrectly classified pixels, and missed pixels, respectively.

Precision emphasizes the purity of the model's predictions, that is, its ability to reduce false detections. Its formula is expressed as:

$$Precision = \frac{TP}{TP + FP} \tag{11}$$

Recall reflects the model's ability to cover positive samples, that is, the completeness of target detection. Its formula is expressed as:

https://doi.org/10.3390/rs18010059

$$Recall = \frac{TP}{TP + FN} \tag{12}$$

The F1 score, as the harmonic mean of Precision and Recall, provides a balanced measure that considers both metrics simultaneously. It prevents overemphasis on high Recall at the expense of Precision, and vice versa. Thus, F1 can achieve a balance between Precision and Recall to a certain extent. Its formula is expressed as:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{13}$$

4.3. Loss Function

In this study, we employ the standard cross-entropy (CE) loss to optimize the binary semantic segmentation model. CE loss is one of the most commonly used loss functions in classification and segmentation tasks, and its core idea is to measure the discrepancy between the predicted probability distribution and the true label distribution. Given a training set containing N samples and C = 2 classes, the cross-entropy loss is defined as follows:

$$Loss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{u=1}^{C} y_{i,u} log(p_{i,u}) \tag{14}$$

here, $y_{i,u}$ denotes the ground truth label of sample i for class u (u = 1, 2), and $p_{i,u}$ represents the probability predicted by the model for sample i belonging to class u after the Softmax function. N is the total number of samples, and C is the total number of classes. This loss function effectively constrains the model's learning process, guiding the predictions to gradually approach the true distribution, thereby improving segmentation accuracy.

4.4. Implementation Details

All experiments in this study were implemented using the PyTorch 2.9.1 framework and executed on a workstation equipped with an NVIDIA GeForce RTX 5080 GPU with 16 GB of memory. The input images were set to a size of 256 × 256 pixels, with a batch size of 16. After multiple trials, a training schedule of 150 epochs was selected to ensure that the model neither overfits nor underfits. The AdamW optimizer was used with an initial learning rate of 0.00006. During training, data augmentation strategies such as random flipping and photometric distortion were applied to enhance the model's generalization capability.

4.5. Experimental Results and Analysis

To demonstrate the effectiveness of our method for semantic segmentation of remote sensing images containing complex landforms such as plateau karst and landslide-prone areas, we compared it with several SOTA segmentation models developed in the past five years, as listed below:

(1) SegNeXt [26]: This network employs a multi-scale convolutional encoder to achieve spatial attention. It maintains relatively low computational complexity while effectively capturing object details, enabling efficient image semantic segmentation.
(2) Swin Transformer [19]: This network utilizes a shifted-window attention mechanism, offering flexibility for multi-scale modeling, thereby enabling effective image semantic segmentation.
(3) SegFormer [20]: By combining a Transformer with lightweight Multi-layer Perceptrons (MLPs), this network integrates local and global attention in a simple, lightweight design to achieve efficient semantic segmentation.
(4) Mask2Former [58]: This network employs a mask-based attention mechanism to extract local features, addressing semantic segmentation tasks across diverse domains.

https://doi.org/10.3390/rs18010059

(5) K-Net [59]: This network uses a set of learnable kernels to achieve consistent segmentation across different semantic categories.
(6) Segmenter [60]: The model performs global context modeling at both the first layer and throughout the network, effectively handling semantic segmentation tasks on small image regions.
(7) QSD [61]: This network introduces a dual-optimization strategy to correct the information distribution in quantized pulse-driven self-attention, enabling efficient performance while maintaining high energy efficiency for semantic segmentation tasks.
(8) LOGCAN++ [3]: The network implements a local–global perception strategy, using local class centers as intermediate perception elements to suppress intra-class differences, thereby achieving higher semantic segmentation accuracy.

4.5.1. Quantitative Analysis

Compared with the mainstream semantic segmentation methods, the segmentation efficiency of LEnet is more outstanding. For instance, compared with the Swin Transformer using the UperNet decoder, our floating-point operations (FLOPS) have decreased by 46.541 G, accounting for only about 22.4% of the total FLOPS of Swin Transformer. At the same time, the number of parameters (Params) has decreased by 10.619 M, accounting for approximately 82% of the total parameters of Swin Transformer. However, the inference speed relative to Swin Transformer is not significantly reduced. Its inference speed still ranks third among all the models we compared. Therefore, it proves that using the strip convolution to simulate the attention feature expert enhancer significantly reduces the computational complexity of long-distance feature modeling. Moreover, the CSA with sparse matrices suppresses irrelevant background noise during the multi-scale feature fusion process, thereby achieving more efficient segmentation for complex landform remote sensing images. The detailed comparison results are shown in Table 1.

Table 1. Compare FLOPs and Params with the models that achieve high performance on our karst dataset. At the same time, we also tested the inference speeds of different models. The input image size is 256 × 256 × 3.

Method	Encoder	Decoder	Flops (G)	Params (M)	FPS (img/s)
SegNeXt	MSCAN-l	Hamhead	16.37	48.775	24.95
SwinUperNet	Swin-t	UperNet	59.97	58.942	33.56
Segformer	Mit-b4	SegHead	13.052	61.37	19.5
Mask2former	Res-50	Mask2former	--	--	16.97
K-Net	Res-50	IterativeDecoder	70.952	79.027	35.5
Segmenter	Vit-b	SegmenterHead	29.903	104.45	13.66
QSD	Spiking-vit	QFPN	16.789	16.519	12.22
LOGCAN++	Swin-t	LOGCANPLUS	28.588	43.15	10.99
LENet(ours)	Swin-t	LENetHead	13.429	48.323	27.04

Compared with Swin Transformer (using the UperNet [25] decoder), the LENet model achieves a 1.3% improvement in IoU, a 0.92% increase in F1 score, and a 1.8% gain in recall on the PKLD karst dataset. This improvement is attributed to LENet's axial semantic modeling mechanism, which captures long-range dependencies overlooked in the UperNet design. Although UperNet leverages multi-scale features through FPN and PPM structures, it does not explicitly focus on key intra-class information during multi-scale feature fusion. LENet addresses this limitation with the Feature Expert Compensator, enabling more complete intra-class feature representation, which is reflected in the notable increase in recall. On the GVLM landslide dataset, LENet achieves a 0.42% improvement in IoU, a 0.26% increase in F1, and a 0.27% gain in recall. Compared with karst geomorphology, landslide

https://doi.org/10.3390/rs18010059

regions exhibit more distinct spectral differences; for example, karst areas may contain partial vegetation coverage, whereas landslide regions rarely have obvious vegetation due to their formation mechanism. This results in generally higher performance metrics for landslide segmentation compared with karst segmentation across different models.

Overall comparison with other SOTA semantic segmentation models shows that LENet achieves excellent performance across comprehensive metrics such as F1, IoU, and mIoU, demonstrating its advancements in complex geomorphology segmentation tasks. It is noteworthy that, as shown in Tables 2 and 3, LENet does not always achieve the highest precision. This is because precision and recall represent a dynamic trade-off, and due to the unique distribution of pixels in complex geomorphological areas, higher recall and F1 scores often correspond to results that more accurately reflect the true complex geomorphological classes. The comparative experiments on the PKLD karst dataset are presented in Table 2, while those on the GVLM landslide dataset are shown in Table 3.

Table 2. The performance was compared with advanced performance models under the karst dataset we produced, with the core metrics being IoU, Recall and F1 score (Unit: %).

Method	Encoder	Decoder	F1	Precision	Recall	IoU	mIOU
SegNeXt	MSCAN-l	Hamhead	82.22	83.77	80.74	69.81	82.07
SwinUperNet	Swin-t	UperNet	81.7	81.87	81.53	69.09	81.56
Segformer	Mit-b4	SegHead	82.22	83.28	81.18	69.81	82.05
Mask2former	Res-50	Mask2former	81.73	82.06	81.4	69.1	81.59
K-Net	Res-50	IterativeDecoder	82.02	81.03	83.05	69.53	81.79
Segmenter	Vit-b	SegmenterHead	80.93	81.95	79.94	67.97	80.93
QSD	Spiking-vit	QFPN	81.59	81.27	81.92	68.91	81.45
LOGCAN++	Swin-t	LOGCANPLUS	81.96	81.16	82.78	69.44	81.75
LENet(ours)	Swin-t	LENetHead	82.62	81.93	83.33	70.39	82.33

Table 3. A comparative experiment was conducted on the public dataset of GVLM decline with Iou, Recall and F1 as the core evaluation metrics (Unit: %).

Method	Encoder	Decoder	F1	Precision	Recall	IoU	mIoU
SegNeXt	MSCAN-l	Hamhead	89.23	88.58	89.9	80.56	88.21
SwinUperNet	Swin-t	UperNet	89.21	87.4	91.11	80.53	88.15
Segformer	Mit-b4	SegHead	88.28	90.3	86.34	79.02	87.5
Mask2former	Res-50	Mask2former	88.93	86.89	91.08	80.07	87.86
K-Net	Res-50	IterativeDecoder	88.79	87.52	90.1	79.83	87.74
Segmenter	Vit-b	SegmenterHead	88.08	88.28	87.87	78.69	87.08
QSD	Spiking-vit	QFPN	89.37	88.46	90.29	80.77	88.23
LOGCAN++	Swin-t	LOGCANPLUS	89.44	88.4	90.51	80.9	88.41
LENet(ours)	Swin-t	LENetHead	89.47	87.65	91.38	80.95	88.42

4.5.2. Qualitative Analysis

As shown in Figures 8 and 9, we selected three representative images from each dataset to compare the segmentation performance of LENet with other models for karst and landslide complex geomorphologies, where GT denotes the Ground Truth. Figure 8 corresponds to the karst dataset, and Figure 9 corresponds to the landslide dataset.

For the first image in Figure 8, there are significant intra-class feature differences, and some boundary information overlaps with shadows, increasing the difficulty of boundary segmentation. Segmenter and K-Net encounter boundary confusion and intra-class inconsistency, respectively. LENet, in contrast, maintains better intra-class integrity and avoids background misclassification issues seen in SegNeXt. In the second image, the

https://doi.org/10.3390/rs18010059

spectral difference between the karst region and its background is small, and the karst regions are distributed sparsely, making it harder for models to learn long-range semantic relationships, which can lead to intra-class information loss. Swin Transformer, Segformer, Mask2former, and QSD all show intra-class misclassification, splitting a single continuous karst region into disconnected parts. LENet, while showing minor intra-class loss in the bottom-right area, successfully preserves most intra-class information and achieves the best intra-class completeness in the larger karst region in the upper-left. In the third image with a complex background, karst regions display noticeable intra-class spectral and texture variations and interference from similar background features. Most segmentation models suffer from some degree of intra-class information loss, whereas LENet produces the most accurate intra-class information consistent with the ground truth.

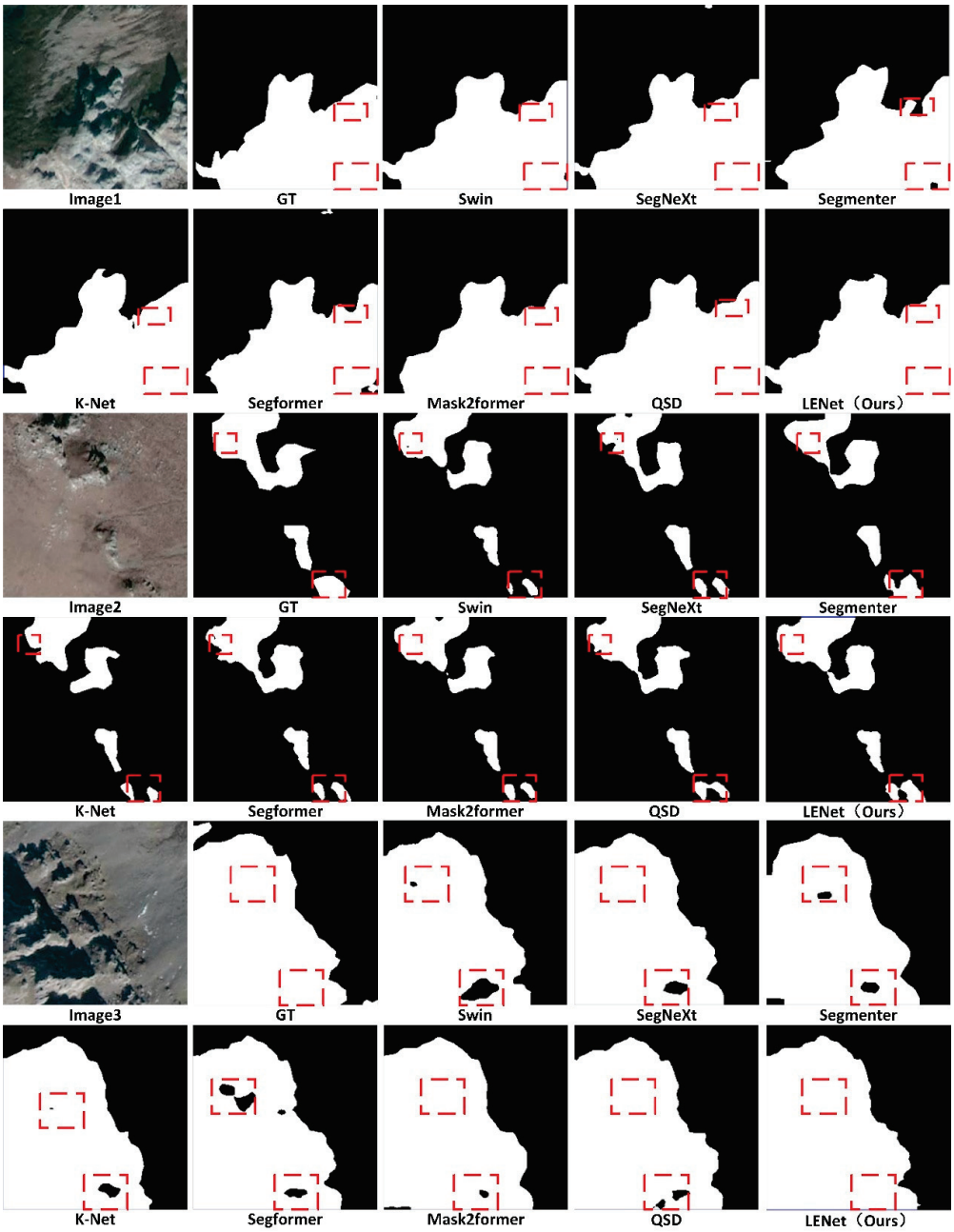


Figure 8. Visualization effect of semantic segmentation results of kast landforms using different models on the validation set (the white part represents the karst area, the black part represents the background, and the red box line represents the key area).

https://doi.org/10.3390/rs18010059

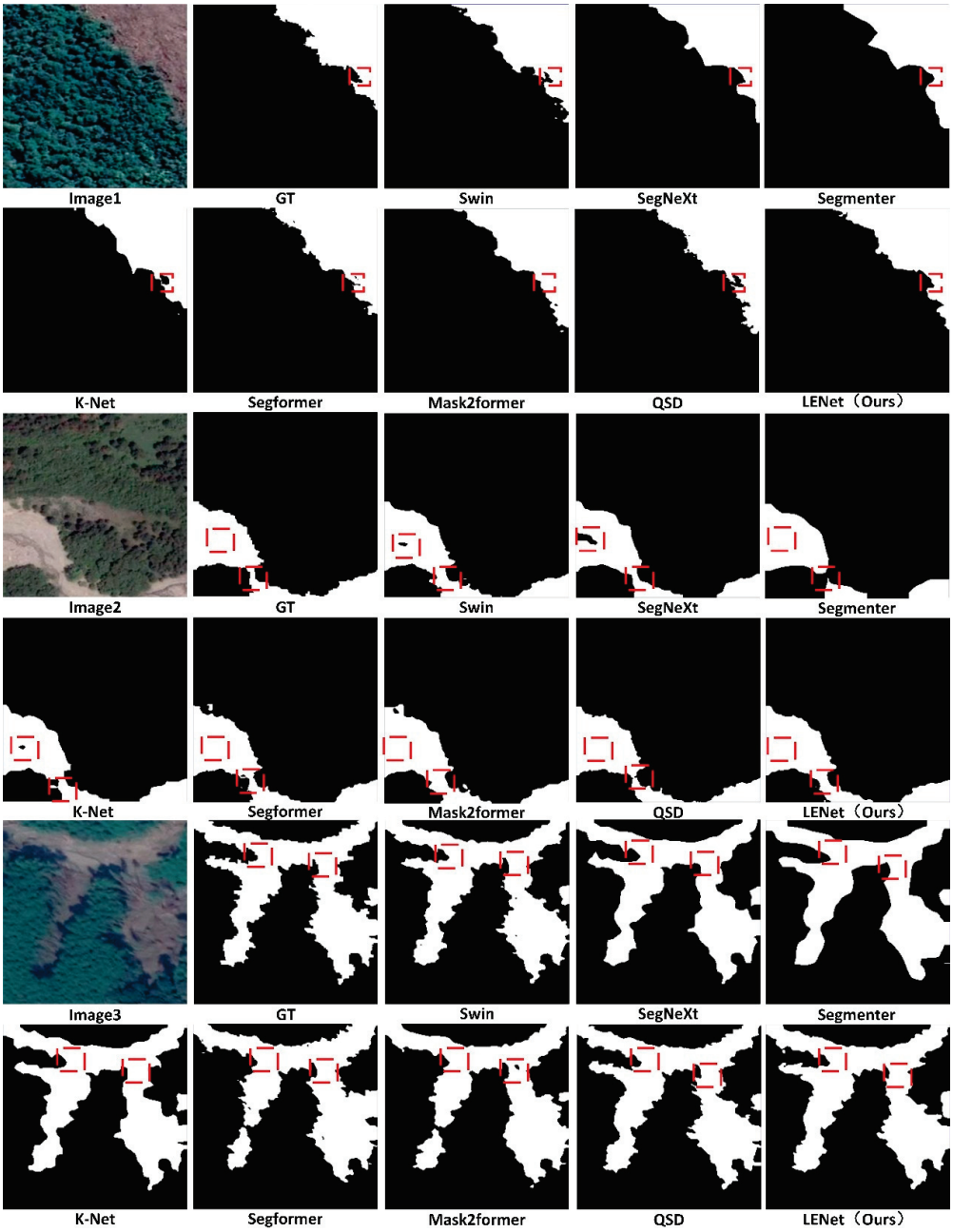


Figure 9. Visualization effect of semantic segmentation results of landslide landforms using different models on the validation set (the white part represents the landslide area, the black part is the background area, and the red box lines are the key areas).

Regarding the landslide segmentation results in Figure 9, the first image shows a landslide region adjacent to vegetated areas. While spectral differences are clear, regions near the connecting boundaries are prone to intra-class misclassification; K-Net and Segformer misclassify parts of the boundary-adjacent areas, whereas LENet preserves intra-class integrity and yields segmentation results closer to the true boundaries. In the second image, the landslide morphology is irregular and connected regions are prone to misclassification, with intra-class spectral variations resembling background ground features. Swin Transformer and SegNeXt misclassify parts of the intra-class regions, and Segmenter shows severe feature loss at the landslide connections. LENet, however, maintains intra-class completeness while retaining local detail at connections. In the third image, some intra-class regions are shadowed, which increases the likelihood of intra-class information loss, and

https://doi.org/10.3390/rs18010059

boundary shapes are highly irregular. Mask2former shows significant intra-class feature loss, resulting in fragmented landslide regions, while SegNeXt, Segmenter, and K-Net produce overly coarse boundaries. LENet, by contrast, preserves intra-class features and produces boundary segmentation results closer to the ground truth.

Overall, these comparative analyses demonstrate that LENet excels in segmenting complex geomorphologies in diverse and complex remote sensing backgrounds, providing superior semantic recognition and visualization performance.

5. Discussion

In this chapter, we will discuss the ablation experiments of the model and the overall discussion of the final results.

5.1. Ablation Study

To further demonstrate the contributions of each component in the LENet to its performance in recognizing complex landscapes, we conducted a series of ablation experiments on the Karst validation dataset. These experiments involved ablation treatments of the entire network architecture, the FEC, the CSA module parameters, and the effective receptive field of the EEBlock. The experimental results are helpful in determining the most effective model design. This section will detail the settings and results of the ablation experiments.

5.1.1. Ablation of the Model Structure

We first conducted ablation experiments on the network architecture to demonstrate the contributions of the FEC, the EEBlock, and the CSA module to the task of complex geomorphology segmentation. Five ablation configurations were designed to evaluate the effectiveness of each component, using IoU, F1, and Recall as evaluation metrics: (1) Using both the FEC and CSA. (2) Using only the Feature Expert Compensator. (3) Using the EEBlock and CSA. (4) Using only the EEBlock (5) Using only the CSA module. The results of these ablation experiments are summarized in Table 4.

Table 4. Ablation experiments on the overall structure of the model (where $\surd$ is used and $\times$ is not used).

EEBlock	FEC	CSA	IoU (%)	F1 (%)	Recall (%)
$\surd$	$\surd$	$\surd$	70.39	82.62	83.33
$\times$	$\surd$	$\surd$	70.03 $\downarrow$ 0.36	82.38 $\downarrow$ 0.24	82.17 $\downarrow$ 1.16
$\times$	$\surd$	$\times$	69.51 $\downarrow$ 0.88	82.01 $\downarrow$ 0.61	81.88 $\downarrow$ 1.45
$\surd$	$\times$	$\surd$	69.63 $\downarrow$ 0.76	82.1 $\downarrow$ 0.52	81.9 $\downarrow$ 1.43
$\surd$	$\times$	$\times$	69.4 $\downarrow$ 0.99	81.94 $\downarrow$ 0.68	81.8 $\downarrow$ 1.53
$\times$	$\times$	$\surd$	69.46 $\downarrow$ 0.93	82.12 $\downarrow$ 0.5	81.95 $\downarrow$ 1.38

Analysis of the above ablation results indicates the following: when only the FEC is retained, the model can still fuse multi-scale features via the intra-class information compensation mechanism, resulting in an IoU higher than that achieved by retaining only the EEBlock or only the CSA module. This improvement arises because the intra-class feature compensation ensures semantic consistency within each category. However, F1 and Recall scores are lower than when either EEBlock or CSA is used alone, due to insufficient global modeling capability and incomplete global feature representation.

When both FEC and CSA are used simultaneously, all evaluation metrics surpass those of the other ablation configurations, demonstrating that the intra-class information compensation and background noise suppression mechanisms are complementary and constitute the core of LENet. Meanwhile, EEBlock, when working together with FEC and CSA, further enhances axial long-range semantic information, enabling more complete intra-class feature modeling.

https://doi.org/10.3390/rs18010059

5.1.2. Ablation of the Feature Expert Compensator

For complex remote sensing images of geomorphologies with large intra-class differences at different scales, sampling stages often encounter significant spectral and texture variations. These variations can lead to abstracted semantic features and difficulty in intra-class recognition, causing intra-class feature information loss, such as inconsistency within classes or misclassification with the background.

We visualized the semantic feature maps learned at the four stages of the backbone network, as shown in Figure 10. As network depth increases, the lost intra-class information fails to receive sufficiently high semantic weight, reducing its correlation with other regions of the same class and resulting in intra-class inconsistency.

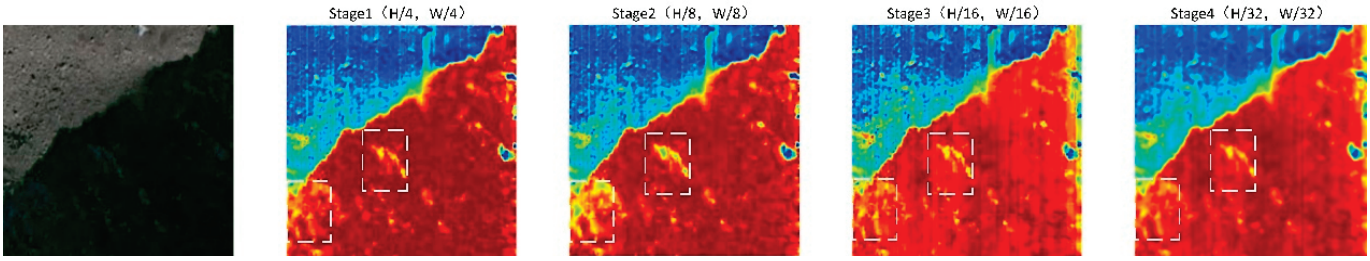


Figure 10. Visualization of feature maps at different stages, where the white box lines represent the key boundary displacement areas, the red part is the part with high class probability, and the blue part is the part with low probability.

Many studies address the loss of fine-grained information by employing residual structures and feature addition to retain detail. While effective for preserving shallow feature details, this approach is insufficient for complex geomorphological remote sensing images where boundary information is intricate and uncertain, and intra-class spectral, texture, and scale variations are pronounced. In such cases, it cannot focus on regions prone to semantic information loss.

To address this, we designed ablation experiments comparing residual addition and feature summation with the FEC that uses learned displacement offsets via DConv. These experiments aim to demonstrate that FEC's ability to learn spatial offsets better compensates for information loss during sampling. We also designed an ablation removing DConv to verify that weighted attention to intra-class information improves segmentation performance. The specific experimental variants are: (1) replacing the FEC with residual addition, and (2) replacing the DConv in FEC with a single-channel summation of max-pooled and average-pooled spatial information. The experimental results are shown in Table 5.

Table 5. Ablation experiment of FEC (where $\surd$ is used and $\times$ is not used).

FEC	Res	DFConv	IoU (%)	F1 (%)	Recall (%)
$\surd$	$\times$	$\surd$	70.39	82.62	83.33
$\times$	$\surd$	$\times$	68.59 $\downarrow$ 1.8	81.37 $\downarrow$ 1.25	81.27 $\downarrow$ 2.06
$\surd$	$\times$	$\times$	69.96 $\downarrow$ 0.43	82.32 $\downarrow$ 0.3	82.12 $\downarrow$ 1.21

Analysis of the ablation experiments on the FEC indicates that its channel–spatial weight separation strategy effectively improves the utilization of multi-scale features, yielding segmentation results closer to the ground truth. Moreover, the deformable convolution in the spatial weight branch, which learns spatial offsets to adaptively weight the convolution kernel, significantly enhances intra-class completeness, as reflected by the noticeable improvement in recall.

https://doi.org/10.3390/rs18010059

5.1.3. Parameter Ablation of Cross-Sparse Attention

In the CSA module, we investigated the impact of different numbers of sparse matrices on modeling and learning the contextual information of spatial offsets. In the experiments, the number of sparse matrix groups was set to 1~5, as shown in Table 6.

Table 6. Ablation experiments of cross sparse attention (Unit: %).

Group Number	IoU	F1	Recall
1	69.08	81.71	81.56
2	69.58	82.06	81.83
3	70.01	82.36	82.59
4	70.39	82.62	83.33
5	70.17	82.36	83.12

The analysis of the above ablation experiments shows that as the number of groups increases, the model's performance first improves and then declines. This is because using too many sparse matrices may cause some relevant features near the weight threshold to be filtered out, and the loss of these features negatively affects the model's segmentation results.

5.1.4. Kernel Size Parameter Effective Receptive Field Ablation Experiments for EEBlock

In EEBlock, we use two banded convolutions to replace the effective receptive field of the large kernel convolution. To demonstrate its effectiveness, the effective receptive fields of standard DWconv, large kernel DWconv, and horizontal and vertical feature extraction strip convolutions are visualized, as shown in Figure 11. And the FLOPs, the number of parameters, and the replacement effect are calculated in Table 7, respectively.

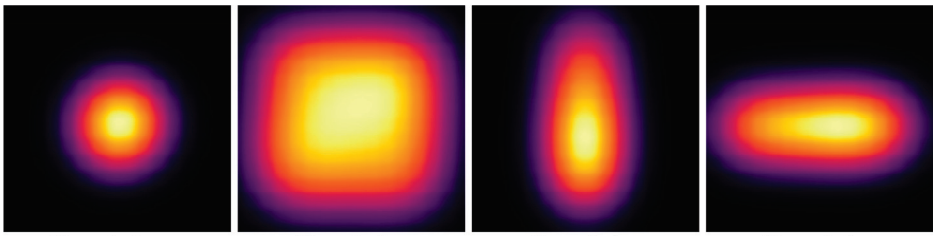


Figure 11. Visualization of effective receptive fields for DWConv with different kernels.

Table 7. Ablation experiments with different kernel parameters and effects of EEBlock.

Kernel Size	FLOPs	Params	IoU (%)	F1 (%)	Recall (%)
(3, 3)	0.11 M	30 B	69.45	81.33	82.07
(11, 11)	1.487 M	366 B	70.35	82.4	83.27
(11, 1) and (1, 11)	0.26 M	72 B	70.39	82.62	83.33

Ablation experimental results and visualization on EEBlock show that the parallel use structure of horizontal and vertical feature extractors can simulate the effective receptive field of a large kernel convolution, and its FLOPs and parameters are significantly lower than those of a large kernel convolution, while achieving the effect of a large kernel convolution. Therefore, the strip convolution feature extraction scheme designed for complex landforms by EEBlock can effectively achieve excellent results while ensuring lightweight.

https://doi.org/10.3390/rs18010059

5.2. Discussion of Final Results

Through the ablation experiments on the model structure and the ablation experiments inside different modules, we can find that LENet is not a single module, but through the overall structure to help LENet for complex landform recognition tasks, which is also the potential advantage of LENet. It is a network with strong internal structure correlation. This strong correlation can help LENet maintain the internal integrity of the categories, which enables it to achieve the best results in the segmentation results.

6. Conclusions

This study explores the challenges faced in the semantic segmentation of complex landform features in remote sensing images, such as significant intra-class scale variations, extremely complex and dynamic information, and difficult-to-model atypical features. To address these issues, a novel semantic segmentation network—LENet—was proposed. This method enhances semantic dependence by performing long-range modeling in both horizontal and vertical directions, introduces a spatial displacement compensation mechanism to strengthen the representation of complex intra-class regions, and combines multi-scale feature fusion to improve the learning of atypical features. Experimental results on the self-built karst landform dataset and the public landslide dataset demonstrate that LENet effectively improves segmentation accuracy, exhibits strong adaptability and robustness. In the future, we will further explore the potential of LENet in other complex landform types and geological disaster scenarios, such as the segmentation task of wind erosion landforms or intelligent identification after earthquakes. We believe that this is a very valuable research method with significant application value, which can expand the depth of research on the intelligent interpretation of remote sensing images.

Author Contributions: Conceptualization, Y.L. and R.Y.; methodology, Y.L. and J.R.; software, Y.L.; validation, J.R., J.W. and S.L.; formal analysis, J.R.; investigation, R.C.; resources, J.R.; data curation, D.Z.; writing—original draft preparation, Y.L.; writing—review and editing, R.Y.; visualization, W.Z. and A.Y.; supervision, J.R.; project administration, J.R.; funding acquisition, J.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Special Project of the Sichuan Bureau of Geology & Mineral Resources, grant number SCDZ-DH202508 and the Academic Degree and Postgraduate Education Reform Project of Sichuan Province, grant number YJGXM24-C073.

Data Availability Statement: Data is available upon request from the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Yuan, Q.; Shen, H.; Li, T.; Li, Z.; Li, S.; Jiang, Y.; Xu, H.; Tan, W.; Yang, Q.; Wang, J.; et al. Deep Learning in Environmental Remote Sensing: Achievements and Challenges. *Remote Sens. Environ.* **2020**, *241*, 111716. [CrossRef]
2. Guo, Y.; Liu, Y.; Georgiou, T.; Lew, M.S. A Review of Semantic Segmentation Using Deep Neural Networks. *Int. J. Multimed. Inf. Retr.* **2018**, *7*, 87–93. [CrossRef]
3. Ma, X.; Lian, R.; Wu, Z.; Guo, H.; Ma, M.; Wu, S.; Du, Z.; Song, S.; Zhang, W. LOGCAN++: Adaptive Local-Global Class-Aware Network for Semantic Segmentation of Remote Sensing Imagery. *arXiv* **2025**, arXiv:2406.16502. [CrossRef]
4. Du, Z.; Yang, J.; Ou, C.; Zhang, T. Smallholder Crop Area Mapped with a Semantic Segmentation Deep Learning Method. *Remote Sens.* **2019**, *11*, 888. [CrossRef]
5. Ma, X.; Lian, R.; Wu, Z.; Guan, R.; Hong, T.; Zhao, M.; Ma, M.; Nie, J.; Du, Z.; Song, S.; et al. A Novel Scene Coupling Semantic Mask Network for Remote Sensing Image Segmentation. *ISPRS J. Photogramm. Remote Sens.* **2025**, *221*, 44–63. [CrossRef]
6. Veress, M. Karst Types and Their Karstification. *J. Earth Sci.* **2020**, *31*, 621–634. [CrossRef]
7. Zhi Gan, Z. Karst Types in China. *GeoJournal* **1980**, *4*, 541–570. [CrossRef]

https://doi.org/10.3390/rs18010059

8. Zhang, X.; Yu, W.; Pun, M.O.; Shi, W. Cross-Domain Landslide Mapping from Large-Scale Remote Sensing Images Using Prototype-Guided Domain-Aware Progressive Representation Learning. *ISPRS J. Photogramm. Remote Sens.* **2023**, *197*, 1–17. [CrossRef]
9. Zhang, Z.; Zhao, T.; Guo, Y.; Yin, J. RS5M and GeoRSCLIP: A Large-Scale Vision- Language Dataset and a Large Vision-Language Model for Remote Sensing. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5642123. [CrossRef]
10. Yu, Z.; Chen, Z.; Xiao, K.; Lei, X.; Tang, R.; He, Q.; Sun, Z.; Guo, H. SegRoadv2: A Hybrid Deformable Self-Attention and Convolutional Network for Road Extraction with Connectivity Structure. *Int. J. Digit. Earth* **2025**, *18*, 2480267. [CrossRef]
11. Semantic Segmentation of Remote Sensing Images: Definition, Methods, Datasets and Applications. In *Communications in Computer and Information Science*; Springer Nature: Cham, Switzerland, 2024; pp. 127–140, ISBN 978-3-031-54320-3.
12. Maulik, U.; Chakraborty, D. Remote Sensing Image Classification: A Survey of Support-Vector-Machine-Based Advanced Techniques. *IEEE Geosci. Remote Sens. Mag.* **2017**, *5*, 33–52. [CrossRef]
13. Kattenborn, T.; Leitloff, J.; Schiefer, F.; Hinz, S. Review on Convolutional Neural Networks (CNN) in Vegetation Remote Sensing. *ISPRS J. Photogramm. Remote Sens.* **2021**, *173*, 24–49. [CrossRef]
14. Lu, W.; Hu, Y.; Shao, W.; Wang, H.; Zhang, Z.; Wang, M. A Multiscale Feature Fusion Enhanced CNN with the Multiscale Channel Attention Mechanism for Efficient Landslide Detection (MS2LandsNet) Using Medium-Resolution Remote Sensing Data. *Int. J. Digit. Earth* **2024**, *17*, 2300731. [CrossRef]
15. Ma, X.; Zhang, X.; Pun, M.-O.; Liu, M. A Multilevel Multimodal Fusion Transformer for Remote Sensing Semantic Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5403215. [CrossRef]
16. Zhang, Y.; Ye, M.; Zhu, G.; Liu, Y.; Guo, P.; Yan, J. FFCA-YOLO for Small Object Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5611215. [CrossRef]
17. Wang, D.; Ma, G.; Zhang, H.; Wang, X.; Zhang, Y. Refined Change Detection in Heterogeneous Low-Resolution Remote Sensing Images for Disaster Emergency Response. *ISPRS J. Photogramm. Remote Sens.* **2025**, *220*, 139–155. [CrossRef]
18. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.
19. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; IEEE: Montreal, QC, Canada, 2021; pp. 9992–10002.
20. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12077–12090.
21. BaoLong, N.; Zhang, C.; Shi, Y.; Hirakawa, T.; Yamashita, T.; Matsui, T.; Fujiyoshi, H. DeBiFormer: Vision Transformer with Deformable Agent Bi-Level Routing Attention. In *Computer Vision—ACCV 2024*; Cho, M., Laptev, I., Tran, D., Yao, A., Zha, H., Eds.; Lecture Notes in Computer Science; Springer Nature: Singapore, 2025; Volume 15481, pp. 445–462, ISBN 978-981-96-0971-0.
22. Wang, L.; Zhang, C.; Meng, X.; Fang, S. A Novel Transformer Based Semantic Segmentation Scheme for Fine-Resolution Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 6506105. [CrossRef]
23. He, X.; Zhou, Y.; Zhao, J.; Zhang, D.; Yao, R.; Xue, Y. Swin Transformer Embedding UNet for Remote Sensing Image Semantic Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 4408715. [CrossRef]
24. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*; Springer: Cham, Switzerland, 2015.
25. Xiao, T.; Liu, Y.; Zhou, B.; Jiang, Y.; Sun, J. Unified Perceptual Parsing for Scene Understanding. In *Computer Vision—ECCV 2018*; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2018; Volume 11209, pp. 432–448, ISBN 978-3-030-01227-4.
26. Guo, M.-H.; Lu, C.-Z.; Hou, Q.; Liu, Z.-N.; Cheng, M.-M.; Hu, S.-M. SegNeXt: Rethinking Convolutional Attention Design for Semantic Segmentation. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 1140–1156.
27. Yu, C.; Wang, J.; Peng, C.; Gao, C.; Yu, G.; Sang, N. BiSeNet: Bilateral Segmentation Network for Real-Time Semantic Segmentation. In *Computer Vision—ECCV 2018*; Springer: Cham, Switzerland, 2018.
28. Zhang, H.; Dana, K.; Shi, J.; Zhang, Z.; Wang, X.; Tyagi, A.; Agrawal, A. Context Encoding for Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
29. Wu, Z.; Ding, T.; Lu, Y.; Pai, D.; Zhang, J.; Wang, W.; Yu, Y.; Ma, Y.; Haeffele, B.D. Token Statistics Transformer: Linear-Time Attention via Variational Rate Reduction. *arXiv* **2024**, arXiv:2412.17810. [CrossRef]
30. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* **2021**, arXiv:2010.11929. [CrossRef]
31. Zhu, L.; Wang, X.; Ke, Z.; Zhang, W.; Lau, R.W.H. BiFormer: Vision Transformer With Bi-Level Routing Attention. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; IEEE/CVF: Vancouver, BC, Canada, 2023; pp. 10323–10333.

https://doi.org/10.3390/rs18010059

32. Fan, Q.; Huang, H.; Chen, M.; Liu, H.; He, R. RMT: Retentive Networks Meet Vision Transformers. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; IEEE: Seattle, WA, USA, 2024; pp. 5641–5651.
33. Zhu, X.; Ma, X.; Su, L.; Jiang, Z.; Du, B.; Wang, X.; Lei, Z.; Feng, W.; Pun, C.-M.; Zhou, J. Mesoscopic Insights: Orchestrating Multi-Scale & Hybrid Architecture for Image Manipulation Localization. In Proceedings of the AAAI Conference on Artificial Intelligence, Philadelphia, PA, USA, 25 February–4 March 2025; Volume 39, pp. 11022–11030.
34. Chen, L.; Fu, Y.; Gu, L.; Yan, C.; Harada, T.; Huang, G. Frequency-Aware Feature Fusion for Dense Image Prediction. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 10763–10780. [CrossRef]
35. Wang, D.; Yang, R.; Zhang, Z.; Liu, H.; Tan, J.; Li, S.; Yang, X.; Wang, X.; Tang, K.; Qiao, Y.; et al. P-Swin: Parallel Swin Transformer Multi-Scale Semantic Segmentation Network for Land Cover Classification. *Comput. Geosci.* **2023**, *175*, 105340. [CrossRef]
36. Rau, J.-Y.; Jhan, J.-P.; Rau, R.-J. Semiautomatic Object-Oriented Landslide Recognition Scheme from Multisensor Optical Imagery and DEM. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 1336–1349. [CrossRef]
37. Cui, J.; Liu, J.; Ni, Y.; Sun, Y.; Guo, M. MCKTNet: Multiscale Cross-Modal Knowledge Transfer Network for Semantic Segmentation of Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 4406015. [CrossRef]
38. He, Q. Prompting Multi-Modal Image Segmentation with Semantic Grouping. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024; Volume 38, pp. 2094–2102. [CrossRef]
39. Zhang, N.; Chai, B.; Song, J.; Tian, T.; Zhu, P.; Ma, J.; Tian, J. Omni-Scene Infrared Vehicle Detection: An Efficient Selective Aggregation Approach and a Unified Benchmark. *ISPRS J. Photogramm. Remote Sens.* **2025**, *223*, 244–260. [CrossRef]
40. Zheng, Z.; Zhong, Y.; Wang, J.; Ma, A.; Zhang, L. FarSeg++: Foreground-Aware Relation Network for Geospatial Object Segmentation in High Spatial Resolution Remote Sensing Imagery. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 13715–13729. [CrossRef]
41. Chen, K.; Liu, C.; Chen, H.; Zhang, H.; Li, W.; Zou, Z.; Shi, Z. RSPrompter: Learning to Prompt for Remote Sensing Instance Segmentation Based on Visual Foundation Model. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 4701117. [CrossRef]
42. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable Convolutional Networks. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
43. Dong, X.; Qin, Y.; Fu, R.; Gao, Y.; Liu, S.; Ye, Y.; Li, B. Multiscale Deformable Attention and Multilevel Features Aggregation for Remote Sensing Object Detection. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 6510405. [CrossRef]
44. Hu, P.-C.; Chen, S.-B.; Huang, L.-L.; Wang, G.-Z.; Tang, J.; Luo, B. Road Extraction by Multiscale Deformable Transformer from Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 2503905. [CrossRef]
45. Şener, A.; Ergen, B. LandslideSegNet: An Effective Deep Learning Network for Landslide Segmentation Using Remote Sensing Imagery. *Earth Sci. Inf.* **2024**, *17*, 3963–3977. [CrossRef]
46. Soares, L.P.; Dias, H.C.; Garcia, G.P.B.; Grohmann, C.H. Landslide Segmentation with Deep Learning: Evaluating Model Generalization in Rainfall-Induced Landslides in Brazil. *Remote Sens.* **2022**, *14*, 2237. [CrossRef]
47. Huang, T.; Huang, B.; Li, S.; Zhao, H.; Yang, X.; Zhu, J. SwinClustering: A New Paradigm for Landscape Character Assessment through Visual Segmentation. *Front. Environ. Sci.* **2025**, *13*, 1509113. [CrossRef]
48. Cheng, J.; Ye, L.; Guo, Y.; Zhang, J.; An, H. Ground Crack Recognition Based on Fully Convolutional Network with Multi-Scale Input. *IEEE Access* **2020**, *8*, 53034–53048. [CrossRef]
49. Goswami, M.; Dey, S.; Mukherjee, A.; Mohanty, S.; Pattnaik, P.K. Convolutional Neural Network Segmentation for Satellite Imagery Data to Identify Landforms Using U-Net Architecture. In *Computational Intelligence in Pattern Recognition*; Springer: Singapore, 2025.
50. Zhou, G.; Chen, W.; Qin, X.; Li, J.; Wang, L. Lithological Unit Classification Based on Geological Knowledge-Guided Deep Learning Framework for Optical Stereo Mapping Satellite Imagery. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5624916. [CrossRef]
51. Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 834–848. [CrossRef]
52. Zhou, S.; Chen, D.; Pan, J.; Shi, J.; Yang, J. Adapt or Perish: Adaptive Sparse Transformer with Attentive Feature Refinement for Image Restoration. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; IEEE: Seattle, WA, USA, 2024; pp. 2952–2963.
53. Wang, S.; Li, B.Z.; Khabsa, M.; Fang, H.; Ma, H. Linformer: Self-Attention with Linear Complexity. *arXiv* **2020**, arXiv:2006.04768. [CrossRef]
54. Zhang, Y.; Zhou, S.; Li, H. Depth Information Assisted Collaborative Mutual Promotion Network for Single Image Dehazing. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; IEEE: Seattle, WA, USA, 2024; pp. 2846–2855.
55. Chen, H.; Chen, X.; Lu, J.; Li, Y. Rethinking Multi-Scale Representations in Deep Deraining Transformer. *AAAI* **2024**, *38*, 1046–1053. [CrossRef]

https://doi.org/10.3390/rs18010059

56. Chen, X.; Li, H.; Li, M.; Pan, J. Learning A Sparse Transformer Network for Effective Image Deraining. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; IEEE: Vancouver, BC, Canada, 2023; pp. 5896–5905.
57. Zhong, D.; Cai, L.; Li, S.; Wang, W.; Zhu, Y.; Liu, Y.; Yang, R. RAP-Net: A Region Affinity Propagation-Guided Semantic Segmentation Network for Plateau Karst Landform Remote Sensing Imagery. *Remote Sens.* **2025**, *17*, 3082. [CrossRef]
58. Cheng, B.; Misra, I.; Schwing, A.G.; Kirillov, A.; Girdhar, R. Masked-Attention Mask Transformer for Universal Image Segmentation. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022.
59. Zhang, W.; Pang, J.; Chen, K.; Loy, C.C. K-Net: Towards Unified Image Segmentation. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 10326–10338.
60. Strudel, R.; Garcia, R.; Laptev, I.; Schmid, C. Segmenter: Transformer for Semantic Segmentation. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; IEEE: Montreal, QC, Canada, 2021; pp. 7242–7252.
61. Qiu, X.; Zhang, M.; Zhang, J.; Wei, W.; Cao, H.; Guo, J.; Zhu, R.-J.; Shan, Y.; Yang, Y.; Li, H. Quantized Spike-Driven Transformer. *arXiv* **2025**, arXiv:2501.13492. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

https://doi.org/10.3390/rs18010059

remote sensing

MDPI

Article

A Multi-Channel Convolutional Neural Network Model for Detecting Active Landslides Using Multi-Source Fusion Images

Jun Wang [1], Hongdong Fan [1,2,*], Wanbing Tuo [3] and Yiru Ren [1]

[1] School of Environment and Spatial Informatics, China University of Mining and Technology, Xuzhou 221116, China; junecattle@cumt.edu.cn (J.W.); ts25160187p31@cumt.edu.cn (Y.R.)
[2] Shanxi Technical Innovation Center for Comprehensive Monitoring and Emergency Disaster Reduction in Mining Areas, Datong 037000, China
[3] School of Engineering, Qinghai Institute of Technology, Xining 810016, China; wbtuo@qhit.edu.cn
* Correspondence: ywfhd@cumt.edu.cn; Tel.: +86-158-621-65361

Highlights

What are the main findings?

- Based on multi-source fusion images, a new dataset and model were constructed for active landslide detection.
- The model introduces a Landslide Attention Module, which has a significant effect on improving the model's performance in detecting active landslides.

What are the implications of the main findings?

- The proposed model achieves superior overall performance and generalization.
- Training with multi-source fusion images enhances performance and efficiency while reducing computation and parameters.

Abstract

Synthetic Aperture Radar Interferometry (InSAR) has demonstrated significant advantages in detecting active landslides. The proliferation of computing technology has enabled the combination of InSAR and deep learning, offering an innovative approach to the automation of landslide detection. However, InSAR-based detection faces two persistent challenges: (1) the difficulty in distinguishing active landslides from other deformation phenomena, which leads to high false alarm rates; and (2) insufficient accuracy in delineating precise landslide boundaries due to low image contrast. The incorporation of multi-source data and multi-branch feature extraction networks can alleviate this issue, yet it inevitably increases computational cost and model complexity. To address these issues, this study first constructs a multi-source fusion image dataset combining optical remote sensing imagery, DEM-derived slope information, and InSAR deformation data. Subsequently, it proposes a multi-channel instance segmentation framework named MCLD R-CNN (Multi-Channel Landslide Detection R-CNN). The proposed network is designed to accept multi-channel inputs and integrates a landslide-focused attention mechanism, which enhances the model's ability to capture landslide-specific features. The experimental findings indicate that the proposed strategy effectively addresses the aforementioned challenges. Moreover, the proposed MCLD R-CNN achieves superior detection accuracy and generalization ability compared to other benchmark models.

Keywords: multi-source fusion images; deep learning; active landslide detection; InSAR; multi-channel convolutional neural network

1. Introduction

Globally, landslides represent one of the most devastating geological disasters, causing considerable risks to human life, infrastructure, and socio-economic stability annually [1]. As human activities continue to intensify, the occurrence frequency and geographic spread of landslides have progressively escalated, resulting in a significant increase in the associated risks to society [2]. Therefore, conducting high-precision landslide detection and mapping their spatial distribution is of great scientific significance. Furthermore, it is of practical importance for the timely implementation of disaster prevention and mitigation measures, ultimately helping to reduce disaster losses [3,4].

Conventional landslide detection methods have primarily relied on field surveys, offering precise and dependable data regarding landslides [5]. However, this approach is time-consuming, labor-intensive, and costly, rendering it unsuitable for large-scale landslide mapping [6]. In contrast, remote sensing technology, known for its extensive spatial coverage and high detection efficiency, has demonstrated significant promise in landslide detection and is extensively used in geological hazard monitoring. For instance, the integration of optical remote sensing and Light Detection and Ranging (LiDAR) techniques has significantly improved landslide detection efficiency [7–9]. Nevertheless, both optical imagery and LiDAR mainly rely on changes in surface morphology or structural features. Consequently, they are effective only for detecting landslides that exhibit substantial surface displacement. Their capability to identify active landslides with slow movements but without noticeable surface changes remains limited. These undetected active landslides may pose potential long-term threats to local residents and infrastructure. Recently, Interferometric Synthetic Aperture Radar (InSAR) technology has advanced significantly. Through interferometric processing of multi-temporal SAR images collected from the same region, InSAR can precisely extract high-resolution data on surface topography and deformation [10–14]. Given InSAR's capability to detect millimeter-level ground displacement, it possesses inherent advantages in identifying active landslides [15]. Compared with other techniques, InSAR provides all-weather, day-and-night, and high-precision monitoring capabilities [16,17], and has thus been extensively applied in active landslide detection studies [18–21]. However, large-scale InSAR-based active landslide detection still faces major challenges due to a lack of efficient, automated interpretation methods. Processing massive interferometric datasets remains computationally complex and heavily dependent on manual intervention. Consequently, the establishment of an automated and large-scale detection framework based on InSAR has emerged as an important research hotspot in recent years.

Driven by the rapid progress of computing technology, deep learning has evolved rapidly, offering an innovative approach for achieving automated landslide detection [22–24]. Several studies have combined optical remote sensing imagery and DEM-derived information [25,26] with deep learning architectures, including object detection, semantic, and instance segmentation networks [27], enabling the automatic identification of landslides [28,29]. Building on these technological advances, numerous studies have investigated the fusion of InSAR observations with deep learning techniques to achieve large-scale automated identification of active landslides. For instance, integrating InSAR-derived deformation maps with networks like Faster R-CNN and Yolo enables the efficient recognition of active landslide distributions. This effectively achieves automated detection at a regional scale [30–32]. Furthermore, deep

https://doi.org/10.3390/rs18010126

learning models based on semantic segmentation, such as U-Net, can extract semantic features of landslides from InSAR deformation fields. This capability enables precise delineation of landslide boundaries and accurate characterization of their spatial extent [33,34]. This approach effectively represents the geographic location and influence zones of active landslides. In summary, the integration of InSAR and deep learning has not only yielded remarkable success in large-scale automated detection of active landslides [35] but also demonstrates considerable promise in improving both the precision and efficiency of landslide detection.

InSAR is capable of detecting various types of slow ground deformation. However, relying solely on a single InSAR product often introduces uncertainties and potential misclassifications in landslide detection. Specifically, large-scale InSAR-based detection faces two critical bottlenecks. First, relying solely on deformation information often leads to high false alarm rates, as non-landslide geological activities frequently exhibit deformation signatures similar to active landslides. Second, accurately delineating landslide boundaries remains difficult, as deformation maps often lack the textural detail required to define precise edges, resulting in coarse segmentation. To improve the robustness of detection outcomes, several studies have developed training datasets through the fusion of various InSAR-derived products, which strengthens deep learning models in capturing the distinctive characteristics of active landslides [35]. Although such approaches have achieved certain improvements in detection performance, they remain limited in distinguishing other geological activities that exhibit deformation patterns similar to those of landslides. Integrating multi-source data is widely considered an effective strategy to address this issue. This strategy involves synergistically combining InSAR deformation measurements, optical remote sensing imagery, and DEM-derived products. Such integration enables the effective differentiation of active landslides from other ground deformations, thereby significantly enhancing detection accuracy and robustness. Several studies have utilized multi-source datasets within deep learning models featuring multi-branch architectures. These approaches have achieved high-precision automated detection of active landslides [25,36]. Nevertheless, these models are typically complex and computationally intensive, often featuring large numbers of parameters. This results in low training efficiency and limits their applicability for large-scale active landslide detection.

To address the aforementioned limitations, this study integrates optical remote sensing imagery, DEM-derived slope data, and InSAR deformation information to construct a multi-source fusion image dataset for active landslide detection. Building upon this foundation, a single-branch multi-channel convolutional neural network (CNN) is developed to enable efficient and automated detection of active landslides, while accurately outlining their spatial extents. The general framework of the research is presented in Figure 1, and the primary aims of this study are summarized as follows:

(1) To develop a comprehensive dataset for active landslide detection by fusing optical remote sensing imagery, DEM- derived slope information, and InSAR-derived deformation data, with the goal of improving detection accuracy and reliability. The constructed dataset contains multi-source fusion images across multiple spatial scales, which enhances the model's ability to generalize in detecting active landslides of diverse sizes.
(2) To propose an active landslide detection model, namely MCLD R-CNN, which supports multi-channel data input and incorporates a Landslide Attention Module to fully exploit the characteristic features of active landslides embedded in the dataset.
(3) To comprehensively assess how multi-source data influence model performance by contrasting the proposed method with traditional deep learning frameworks and various dataset types, and to further examine the strengths and weaknesses of the model regarding detection accuracy and computational efficiency.

https://doi.org/10.3390/rs18010126

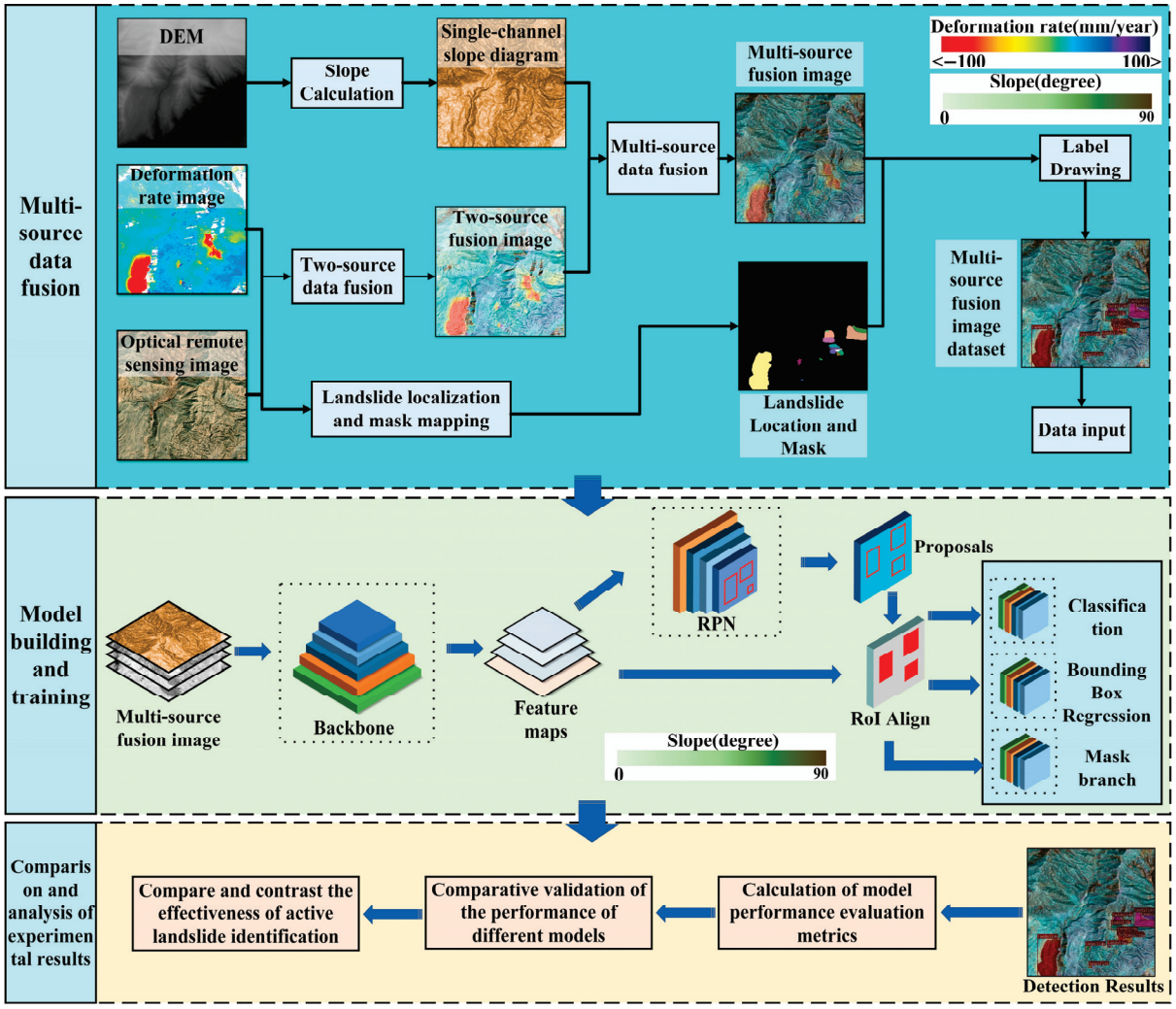


Figure 1. Experimental Workflow.

2. Research Area and Data Sources

2.1. Research Area

The research region is situated in the upper reaches of the Yellow River within Qinghai Province, China (Figure 2). This region encompasses two major geomorphic terraces in China and functions as a transitional zone between the highly uplifted Tibetan Plateau and the comparatively lower Loess Plateau [37]. Due to prolonged northward compression from the Tibetan Plateau, the region has developed a landscape of mountains and basins. Additionally, erosion from the Yellow River and its tributaries has generated geomorphic conditions conducive to landslide formation [38,39]. The region exhibits an arid to semi-arid continental plateau climate, with most of the 316–436 mm annual precipitation falling in summer and autumn. The unique climate and precipitation characteristics contribute to rock weathering and sparse vegetation, further promoting landslide occurrences [40]. Collectively, these factors contribute to the frequent occurrence of geological hazards, including landslides, which are typically large in scale, widely distributed, formed through complex mechanisms, and possess significant destructive potential [41]. Consequently, these landslide activities provide abundant experimental data for this study and serve as an ideal setting for validating the effectiveness of the multi-source fusion image-based active landslide detection method.

https://doi.org/10.3390/rs18010126

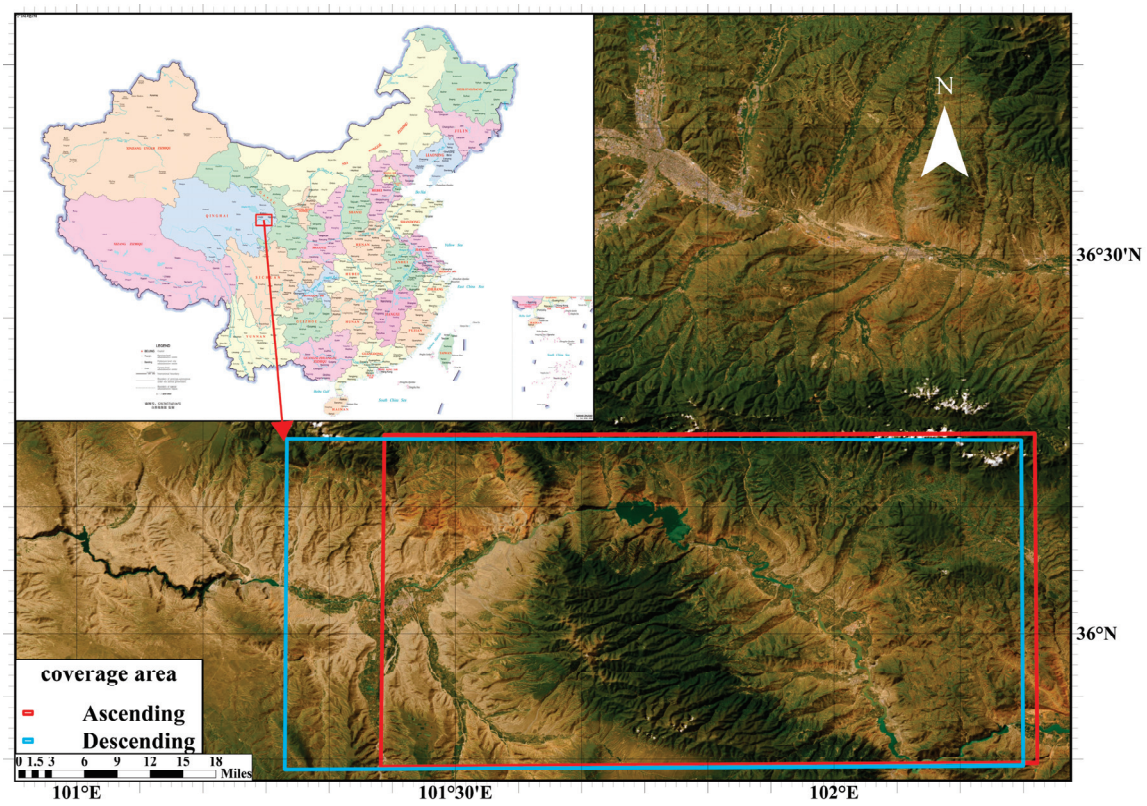


Figure 2. Research region. The red and blue rectangles represent the ascending and descending track image coverage areas, respectively.

2.2. Data Sources

A total of 165 Sentinel-1 SLC images spanning the research region between 2021 and 2023 were selected, as illustrated in Figure 2. As presented in Table 1, the satellite data were collected in IW swath mode featuring HH + VH polarization. The images are divided into 45 ascending tracks and 120 descending track datasets. The ascending track data is located at path 128, frame 114, while the descending track data is located at path 33, frame 472, and path 135, frame 473. High-resolution DEMs obtained from ALOS PALSAR, SRTM datasets, and optical imagery sourced from Google Earth that encompass the study region were also acquired concurrently. The SRTM DEM served as auxiliary data for InSAR processing, whereas the SLC imagery, ALOS-derived DEM, and optical data were integrated to construct a multi-source fusion image dataset aimed at detecting active landslides.

Table 1. Fundamental Specifications of the Utilized Sentinel-1 Imagery.

Beam Modes	Polarizations	Flight Directions	Path	Frame	Number of Images
IW	VV + VH	ASCENDING	128	114	45
IW	VV + VH	DESCENDING	33	472	63
IW	VV + VH	DESCENDING	135	473	57

3. Method

3.1. Data Preparation and Preprocessing

3.1.1. Initial Data Processing

This study utilized the SBAS-InSAR [42] approach, implemented via the GAMMA software (version 2017) suite [43], to process multiple sets of ascending and descending Sentinel-1 SAR data, thereby deriving the temporal ground deformation rates across

https://doi.org/10.3390/rs18010126

the study region. The processing workflow proceeded as follows: First, temporal and spatial baseline thresholds were established (temporal baseline $\leq$ 36 days, perpendicular baseline $\leq$ 200 m) to filter valid interferometric pairs, ensuring the continuity and spatial correlation of the time series. Subsequently, differential processing was performed for each pair with a multi-looking factor of 5 (range) by 1 (azimuth), and high-coherence regions (coherence > 0.3) were selected for phase unwrapping using the Minimum Cost Flow (MCF) algorithm. Then, the Singular Value Decomposition (SVD) method was utilized to solve the SBAS linear equations. Subsequently, spatio-temporal filtering was applied to eliminate or mitigate atmospheric disturbances, residual orbital errors, and noise. Specifically, a spatial low-pass filter with a window size of 256 pixels was applied to estimate and remove the spatially correlated atmospheric phase components. In addition, a temporal filter with a window of 70 days was employed to further reduce temporal noise. This process ultimately yielded high-accuracy surface deformation velocities. Finally, the data were transformed from the radar coordinate system to the WGS84 geographic coordinate system, generating final deformation velocity maps in GeoTIFF format. It is important to note that the deformation results from ascending and descending tracks were maintained as independent Line-of-Sight (LOS) velocity maps for subsequent analysis, rather than being decomposed into vector components. The selection of SBAS-InSAR is justified by its superior monitoring accuracy compared to D-InSAR [44]. Furthermore, unlike PS-InSAR, the SBAS method effectively mitigates interferometric decorrelation, rendering it more suitable for detecting active landslides and characterizing their displacement rates [20,45].

In addition, since the study area covers a large region, multiple ALOS PALSAR scenes were utilized to generate the DEM data. Accordingly, the DEM raster tiles were first mosaicked and merged to produce a continuous elevation surface. The topographic gradient was then calculated to extract slope information. This processed slope map served as the topographic input feature for subsequent multi-source data fusion.

3.1.2. Multi-Source Data Fusion

First, the deformation velocity map, slope map, and Google Earth optical imagery were all projected into the WGS84 geographic coordinate system. Based on their spatial reference information, precise image registration was performed among the three datasets to ensure spatial consistency and the reliability of the multi-source fusion results.

To construct a four-channel input tensor incorporating deformation, texture, and topographic features, we adopted a step-by-step fusion strategy. First, the raw InSAR deformation rate data, denoted as V_{raw} (unit: mm/year), is processed. To suppress the impact of extreme outliers and conform the data to the image format, the deformation rates are clipped within the range $[v_{min}, v_{max}]$ (set to -100 and 100 mm/year in this study) and linearly normalized to obtain $V_{norm} \in [0, 1]$ using the following equation:

$$V_{norm} = \frac{clip(V_{raw}, v_{min}, v_{max}) - v_{min}}{v_{max} - v_{min}} \quad (1)$$

Subsequently, a pseudo-color mapping function is applied to convert the normalized single-channel data into a three-channel RGB deformation image:

$$C_{def} = \Phi(V_{norm}) \quad (2)$$

In the formula, Φ is a pseudo-color mapping function, where C_{def} consists of three components: R_{def}, G_{def}, B_{def}.

Following the generation of the deformation image, to highlight deformation characteristics while preserving surface textures, the optical remote sensing image (C_{opt}) is

https://doi.org/10.3390/rs18010126

fused with C_{def} through a pixel-wise weighted combination. The fused RGB image, C_{fused}, is calculated as:

$$C_{fused} = \lambda \times C_{def} + (1 - \lambda) \times C_{opt}, \lambda \in [0, 1] \quad (3)$$

In the formula, λ represents the fusion weight coefficient. C_{fused} is the two-source fusion image, C_{def} is the deformation rate image, and C_{opt} is the optical remote sensing image.

Finally, to incorporate topographic features, slope data is embedded as the fourth channel (the Alpha channel) of the image. Let S denote the slope value (in degrees) derived from the DEM. We map the slope range of 0° to 90° linearly to the integer range of 0 to 255 to generate the Alpha channel value A:

$$A = round(\frac{S}{90} \times 255) \quad (4)$$

The normalized slope map A is then concatenated with the fused RGB image C_{fused} to generate the final four-channel input. In this configuration, the normalized slope values serve as the Alpha channel (spanning from 0 to 255), effectively acting as a transparency mask. This linear mapping naturally enhances feature distinctiveness. Specifically, regions with high slopes are rendered with higher opacity to highlight terrain undulations. Conversely, low-slope areas remain more transparent to prioritize the underlying deformation features. Consequently, the resulting four-channel multi-source fusion image efficiently integrates deformation, optical, and topographic information within a single tensor. This approach not only enriches the input features but also enhances both the training efficiency and processing speed of the developed model. The complete data fusion process is depicted in Figure 3.

3.2. Dataset Creation

To augment the model's generalization capability for landslides of varying sizes, this research constructed the dataset using multi-source fusion images at five distinct spatial scales, spanning from 1:24,000 to 1:100,000. The imagery was divided into training tiles with a resolution of 512 × 512 pixels. To ensure the validity and robustness of the dataset, this study established an operational definition for "active landslides" that prioritizes InSAR observation while using optical data for auxiliary verification. Considering that slow-moving landslides [35] (typical velocities less than 1.6 m/year) often lack significant surface changes visible in optical imagery, we primarily identified "activity" based on SBAS-InSAR results. Specifically, areas exhibiting continuous ground deformation with an absolute annual velocity exceeding 10 mm/year [46,47] were classified as active candidates.

Building on this definition, active landslide samples were determined by combining existing landslide data [38,41], optical remote sensing imagery and InSAR deformation information. Quantitative statistics based on spatial overlap analysis reveal that approximately 60% of the labeled samples originated from existing inventories and were subsequently refined using InSAR deformation boundaries. The remaining 40% were newly detected primarily through InSAR evidence, characterized by distinct deformation signals in areas previously unmapped or lacking clear optical signs. Existing inventories functioned as spatial priors; however, it is important to note that we did not directly inherit the boundaries from historical records; instead, all targets were re-verified, and their contours were rigorously re-delineated based on the current InSAR deformation fields. Furthermore, considering the spatial resolution of Sentinel-1 data and the reliability of feature extraction, we set the minimum mapping area to 1000 m^2. Deformation areas smaller than this scale were excluded from the repository.

https://doi.org/10.3390/rs18010126

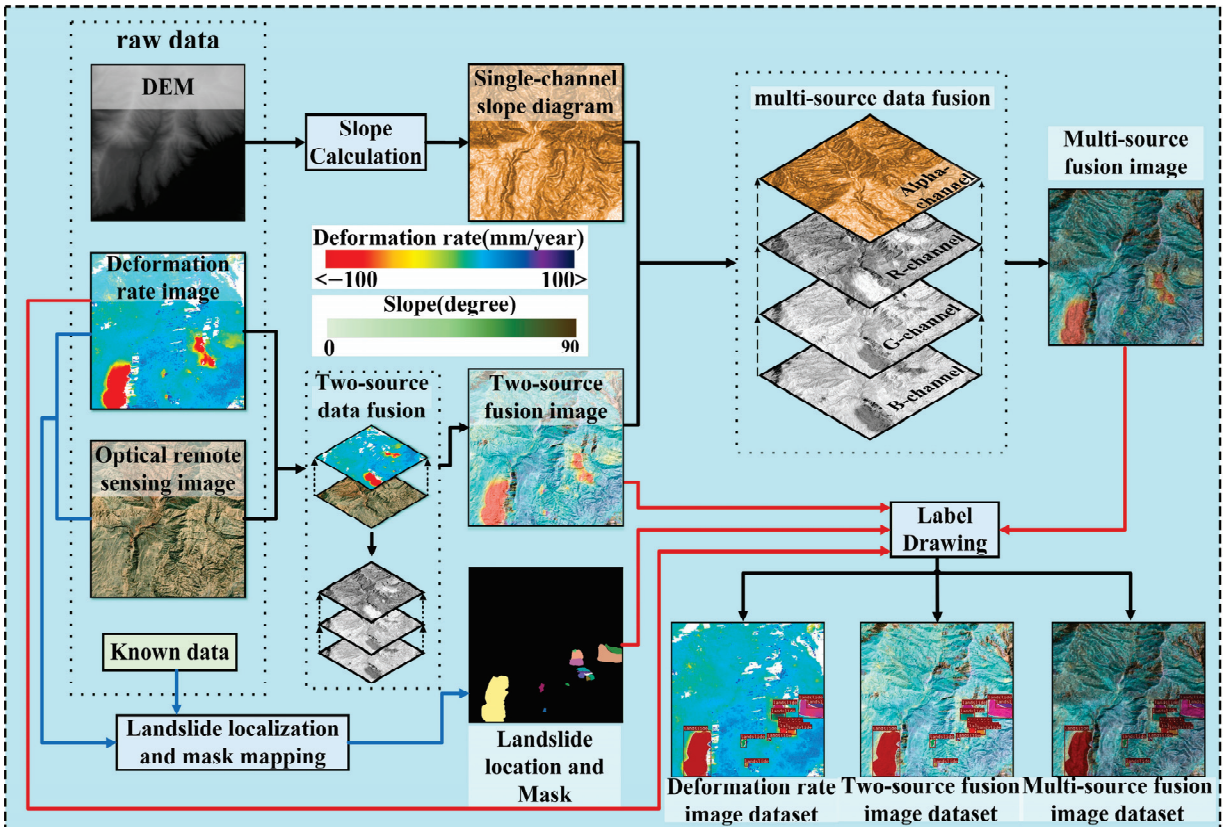


Figure 3. Multi-source fusion image processing flow. The black arrows represent the fusion process, the blue arrows indicate landslide target localization and mask drawing, and the red arrows represent the dataset label creation.

As presented in Figure 4, a total of 437 active landslides were detected from the ascending track imagery and 427 from the descending track imagery. To characterize the dataset and assess potential biases, we analyzed the statistical distribution of the labeled active landslides. The landslide areas exhibit a wide range, spanning from a minimum of 1070 m^2 to a maximum of approximately 6.08 km^2, with the vast majority concentrated between 2000 m^2 and 60,000 m^2. Topographically, the landslides are predominantly located on slopes ranging from 15° to 40°, which aligns with regional landslide susceptibility characteristics reported in [41]. In terms of aspect, the distribution shows a clear preference for east- and south-facing slopes, likely attributable to local microclimatic conditions.

Image augmentation techniques were applied to the sliced images for data expansion. The LabelMe tool was used to manually label the active landslides in the images in COCO dataset format [48]. Following data filtering, the finalized active landslide detection dataset comprised 3450 samples, of which 3105 were allocated for model training and 345 were reserved as an independent test set for performance evaluation. In addition, to evaluate the advantages of multi-source data fusion, a two-source fusion image dataset (fusion of optical remote sensing images and deformation rate images) and a deformation rate image dataset were also created, as illustrated in the bottom-right section of Figure 3. All the aforementioned datasets employed identical mask annotations to maintain consistency and ensure comparability in the training outcomes. An example of the dataset is illustrated in Figure 5.

https://doi.org/10.3390/rs18010126

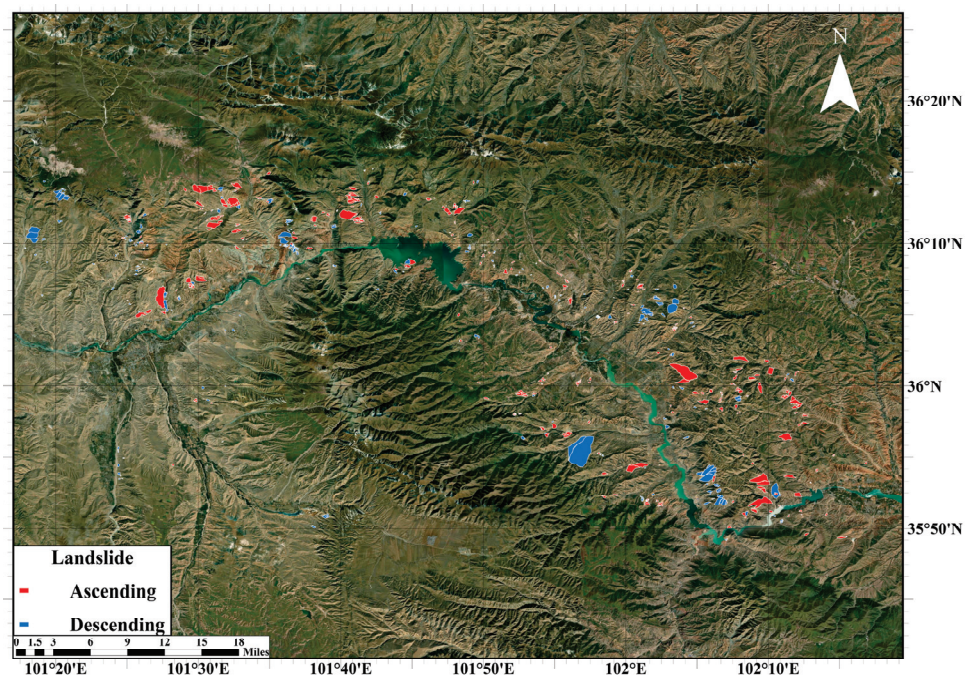


Figure 4. Spatial distribution of active landslides within the study region. Red and blue boxes denote landslides detected from ascending and descending track imagery, respectively.

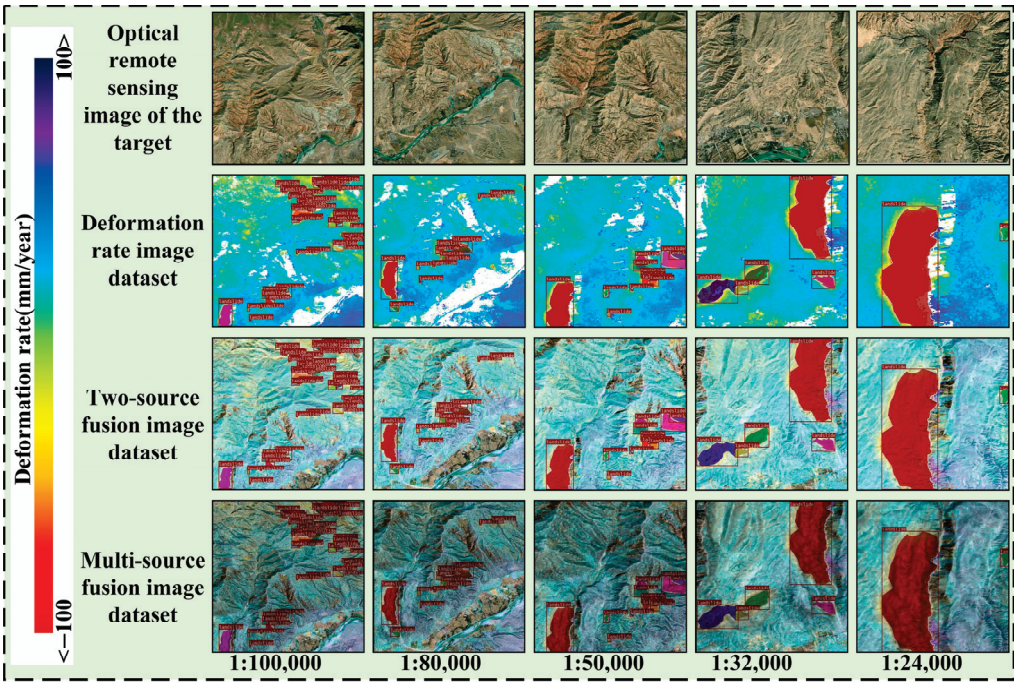


Figure 5. Examples from each dataset.

3.3. MCLD R-CNN

In this research, an instance segmentation model with multi-channel input, termed MCLD R-CNN (Multi-Channel Landslide Detection R-CNN), is constructed upon the Mask R-CNN architecture [49]. MCLD R-CNN functions as a two-stage framework for instance segmentation [50,51]. It is capable of performing both object detection and mask generation simultaneously, thereby enabling precise identification and contour extraction of active landslide areas. The model accommodates multi-channel data input to fully exploit the rich information contained in multi-source fusion imagery. Moreover, a Landslide Attention

https://doi.org/10.3390/rs18010126

Module was embedded within the network architecture to strengthen the model's capacity for extracting and identifying active landslide features. As illustrated in Figure 6a, the MCLD R-CNN primarily consists of a backbone network, a Region Proposal Network (RPN), an RoI Align layer, a detection head, and a mask branch. The following section focuses on the structural improvements incorporated into the proposed model.

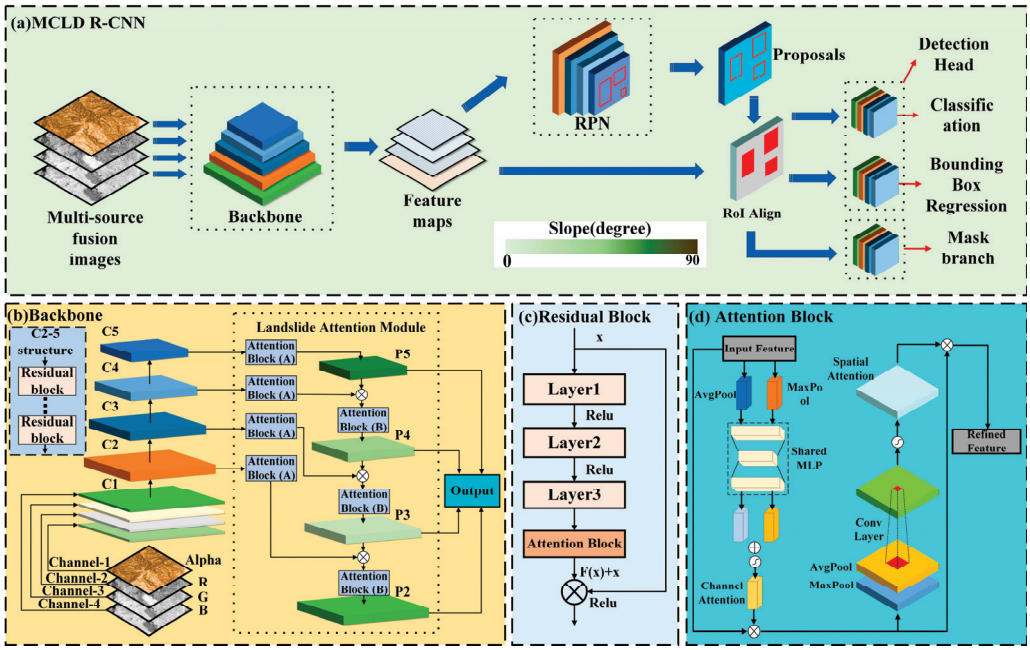


Figure 6. MCLD R-CNN. (**a**) illustrates the overall architecture of the MCLD R-CNN model. (**b**) illustrates the backbone network structure used for extracting target features. (**c**) depicts the structure of the improved residual block that makes up the backbone network. (**d**) demonstrates the attention block used to build the backbone network.

3.3.1. Backbone Network of MCLD R-CNN

The backbone network mainly serves to capture image features at multiple scales from the given input samples. As shown in Figure 6b, it mainly comprises an improved Residual Network (ResNet) [52] and a Landslide Attention Module. The former supports multi-channel data input and is capable of extracting hierarchical features from the input. The latter aims to integrate multi-level features while maintaining both semantic information from higher layers and detailed characteristics from lower ones.

As the depth of convolutional neural networks increases, issues such as vanishing and exploding gradients may arise, impeding the efficiency of model training. ResNet addresses these issues by introducing residual learning and shortcut connections, enabling the stable training of deep networks while significantly enhancing feature extraction capabilities. The core component of the ResNet architecture is the residual block, and by stacking multiple residual blocks, ResNet networks of varying depths can be constructed. The residual block is defined mathematically as follows:

$$b = Y(a, \{K_i\}) + K_s a \tag{5}$$

In this equation, a denotes the input to the residual block, Y (a, $\{K_i\}$) represents the residual function, and K_S is a 1×1 convolution kernel employed to align the dimensions of the input a with the output.

In this work, the ResNet training architecture is adapted to handle multi-channel input data. Additionally, attention blocks (Figure 6c) are integrated into the residual blocks to improve the network's ability to extract features associated with active landslides [53]. The

https://doi.org/10.3390/rs18010126

fundamental structure of the attention block, shown in Figure 6d, comprises a channel attention component and a spatial attention component. The channel attention component assigns importance weights to each channel, thereby strengthening the feature responses of the most relevant channels. In contrast, the spatial attention component computes importance weights for each spatial position, enabling the model to concentrate on the target regions. The fundamental computational formulas for channel and spatial attention are given below:

$$C_{cha}(Y) = Sig(MLP(AP(Y)) + MLP(MP(Y))) \tag{6}$$

$$Y_{cha} = C_{cha}(Y) \otimes Y \tag{7}$$

$$C_{spa}(Y_{cha}) = Sig(Conv^{7\times7}([AP(Y_{cha}); MP(Y_{cha})])) \tag{8}$$

$$Y_{out} = C_{spa}(Y_{cha}) \otimes Y_{cha} \tag{9}$$

In the formula, C_{cha} and C_{spa} denote the channel and spatial attention components, respectively. Y is the input feature map, Y_{cha} is the feature map after channel attention, and Y_{out} represents the final output. $AP(Y)$ and $MP(Y)$ perform average and max pooling over the spatial dimensions $H \times W$. *MLP* denotes a shared two-layer fully connected network. $Conv^{7\times7}$ denotes a 7×7 convolution, and *Sig* is the Sigmoid function that normalizes weights to [0, 1].

The original ResNet architecture outputs only the highest-level features, which are highly abstract and often lack detailed information about target objects. This constraint limits the model's capacity to precisely detect active landslides with diverse shapes and scales. To overcome this limitation, this study proposes a Landslide Attention Module that combines attention blocks with a Feature Pyramid Network (FPN) structure [54]. As illustrated in Figure 6b, the Landslide Attention Module initially extracts multi-level features via lateral connections and refines the original features using Attention Block A. Then, starting from the topmost feature level, a top-down pathway with upsampling is employed to progressively fuse the higher-level features with those from lower levels. Attention block B is applied to reinforce the fused features. This design facilitates the flow of high-level semantic information to lower layers while concurrently strengthening features relevant to the target. In contrast to relying solely on a single high-level feature map, integrating multi-level features retains both semantic and fine-grained information from the input. This integration enhances the model's ability to detect and localize objects, thereby improving the accuracy of both detection and mask generation.

3.3.2. The Improved RPN and the RoI Align Layer

The RPN produces candidate regions of interest, represented as bounding boxes that potentially contain objects, using the feature maps generated by the backbone network. Its core task is to determine whether an anchor box contains an object. Additionally, it adjusts the anchor box's position and size through a regression branch, bringing it closer to the true object bounding box. Finally, high-quality candidate regions are selected through Non-Maximum Suppression.

The specific process is as follows:

(1) Multiple rectangular anchor boxes with varying scales and aspect ratios are predefined for each pixel in the input feature map. For the pixel located at (i, j) in the feature map, the center coordinates of its corresponding anchor boxes are defined as:

$$x_c = (i + 0.5) \times s \tag{10}$$

$$y_c = (j + 0.5) \times s \tag{11}$$

In the formula, s represents the total stride of the backbone network.

https://doi.org/10.3390/rs18010126

(2) Through convolutional processing in the RPN head, the network outputs the objectness score for each anchor box and the bounding box offsets, which determine the candidate regions for subsequent processing. The RPN employs a joint loss function that includes both classification and regression components:

$$L(\{p_i\},\{t_i\})_{RPN} = \frac{1}{N_{cls}}\sum_i L_{cls}(p_i, p_i^*) + \lambda\frac{1}{N_{reg}}\sum_i p_i^* L_{reg}(t_i, t_i^*) \quad (12)$$

In the formula, L_{cls} denotes the binary cross-entropy term, while L_{reg} corresponds to the smooth L1 regression loss. p_i indicates the predicted probability for the foreground class, and $p_i^* \in \{0, 1\}$ is the corresponding ground-truth label. t_i and t_i^* represent the predicted and reference bounding-box offsets, respectively. N_{cls} and N_{reg} serve as normalization terms, and λ acts as a weighting coefficient to balance the two loss components.

The RoI Align layer addresses the key issue of feature alignment for candidate regions through precise coordinate mapping and feature sampling mechanisms. The primary role of this module is to precisely project the Region of Interest (RoI) coordinates from the original image domain onto the feature map domain. It then extracts the corresponding target regions from the feature maps generated by the backbone network. By leveraging bilinear interpolation, RoI Align preserves sub-pixel spatial information during feature sampling. This effectively avoids the feature misalignment problems caused by the double quantization in the traditional RoI Pool method [55]. As a result, it significantly improves spatial alignment accuracy and enhances both the smoothness of pixel-level mask boundaries and the precision of object localization.

As shown in Figure 7, incorporating attention blocks into both the RPN and RoI Align layers optimizes proposal matching and region selection. This enhancement improves the quality of candidate bounding boxes from the RPN and strengthens feature alignment in the RoI Align layer. Consequently, it yields more precise feature representations, enhancing subsequent detection and segmentation performance.

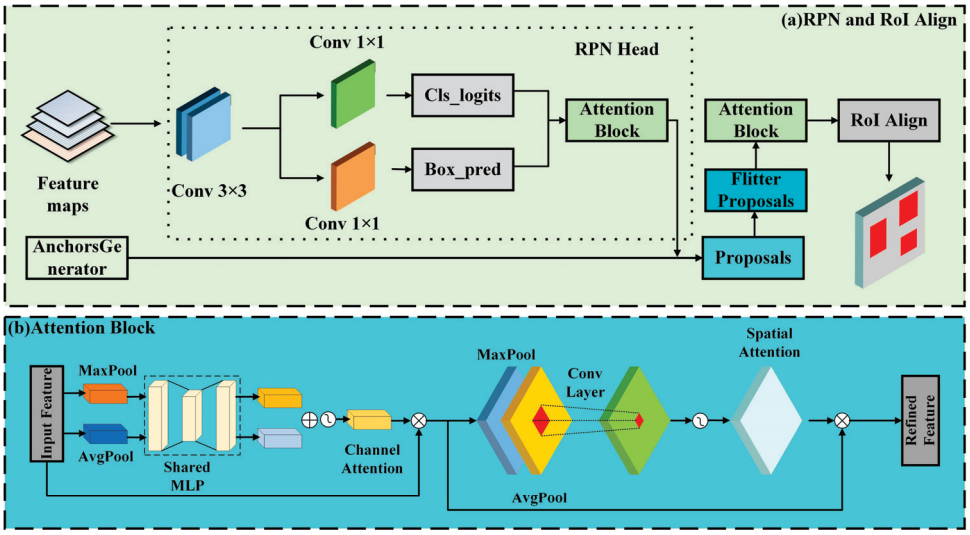


Figure 7. Improved RoI Align Layer and RPN. (**a**) illustrates the framework composition of the RoI Align layers and RPN. (**b**) depicts the attention block incorporated into the framework.

3.4. Model Performance Evaluation Metrics

Five evaluation metrics were employed to assess model performance. The selected metrics include Precision (P), Recall (R), F1-score, mean Average Precision at IoU = 0.5 (mAP_{50}), and mean Average Precision computed over IoU thresholds from 0.5 to 0.95 ($mAP_{50–95}$). Among these metrics, P measures the ratio of correctly identified positive samples to the

https://doi.org/10.3390/rs18010126

total number of samples predicted as positive by the model. R measures the proportion of actual positive samples that are correctly detected by the model. mAP_{50} reflects object detection robustness and coarse localization, whereas $mAP_{50–95}$ evaluates the quality of high-precision bounding box regression and spatial fitting. The F1-score is employed to assess the harmonic balance between precision and recall, providing an integrated measure of the model's overall performance. The computational formulations of these metrics are presented in Equations (10)–(13).

$$P = TP/(FP + TP) \tag{13}$$

$$R = TP/(FN + TP) \tag{14}$$

$$mAP = \frac{1}{c}\sum_{i=1}^{c}\left(\int_0^1 P_i(R)dR\right) \tag{15}$$

$$F1 = 2 \times P \times R/(R + P) \tag{16}$$

In the above equations, *TP* denotes the number of true positive samples correctly identified as positive by the model, *FP* refers to false positive samples that are incorrectly classified as positive, and *FN* indicates false negative samples that are mistakenly predicted as negative. The variable *c* represents the total number of classes.

This study evaluates the performance of the MCLD R-CNN model from two perspectives: active landslide detection and segmentation. Detection performance assesses the model's capability to accurately identify and localize active landslides, while segmentation performance assesses its capability to extract and refine object contours. Both aspects are evaluated using the five aforementioned performance metrics.

4. Results and Analysis of Experiments

4.1. Model Training Parameters

The experiments were conducted in a workstation environment using Python 3.12 as the programming language and PyTorch 1.13 as the deep learning framework. The workstation was equipped with an Intel i5-14600KF CPU and an NVIDIA GeForce RTX 4090D GPU, providing the computational resources necessary for model training and evaluation. The model was trained for 200 epochs with a batch size of 8 and a learning rate of 0.001. To comprehensively assess the performance of MCLD R-CNN, several instance segmentation models were included for comparative evaluation. To ensure the fairness and validity of the comparative experiments, a strict protocol covering both model architecture and training strategies was established. First, regarding architecture adaptation, the input layers of all baseline models were adapted to match the channel dimensions of the dataset. Specifically, for the multi-source fusion image dataset, the first convolutional layers were adjusted to accept four-channel inputs, ensuring that all models had equal access to the slope information. Second, regarding the training process, although hyperparameters such as epoch numbers and learning rates varied to accommodate the distinct convergence characteristics of one-stage (Yolo series) and two-stage (R-CNN series) architectures, the optimization objective remained consistent. Training was conducted for sufficient fixed epochs (as detailed in Table 2) to ensure full model convergence, which was assessed by monitoring the stabilization of validation loss and mAP curves. Crucially, the "best model" checkpoint for each method was selected based on the highest mAP_{50} score achieved on the validation set during training, rather than relying on the final epoch weights. Furthermore, to eliminate data-related bias, an identical data augmentation pipeline was applied

https://doi.org/10.3390/rs18010126

uniformly across all models. The models along with their respective training parameters are summarized in Table 2. For the Mask R-CNN variants, the designations ranging from 50 to 152 represent the varying depths of the backbone networks, corresponding to ResNet architectures with 50 to 152 layers.

Table 2. Training Configurations of Comparative Models.

Model	Epoch	Learning-Rate	Batch Size
Yolov8 [56]	1000	0.001	8
Yolov9 [57]	1000	0.001	8
Yolov11 [58]	1000	0.001	8
Yolov12 [59]	1000	0.001	8
Mask R-CNN-50 [49]	40	0.02	8
Mask R-CNN-101	40	0.02	8
Mask R-CNN-152	40	0.02	8

4.2. Model Performance Comparison

4.2.1. Comparison of Performance Evaluation Metrics

Table 3 provides a summary of the detection and segmentation performance of each model across the three datasets. As shown in the tables:

(1) For the same model across different datasets, the models trained with multi-source fusion images consistently show the best performance, while the models trained with deformation rate image datasets perform inferiorly. This indicates that training the model with fusion images yields better results, and as the variety of fused data increases, the model's performance improves progressively. This is particularly evident in the $mAP_{50–95}$ metric, where the maximum improvements reached 45% (detection performance) and 60% (segmentation performance). Moreover, the Precision (detection performance) of our model improved from 93.95% on the deformation-only dataset to 97.79% on the multi-source dataset. This increase of nearly 4 percentage points signifies a substantial reduction in False Positives (FP), quantitatively confirming that fusing multi-source data effectively filters out confounding geological activities.

(2) Across different models on the same dataset, MCLD R-CNN consistently demonstrates superior performance. Particularly in terms of R, F1, and mAP_{50} metrics, it achieved the highest detection and segmentation scores across all datasets. In some datasets, it also achieved the highest P score, and the $mAP_{50–95}$ metric was only slightly lower than Yolov9. Therefore, based on these comparative results, it is evident that MCLD R-CNN offers the best overall performance.

Table 3. Quantitative Indicators for Each Model Under Different Data Sets.

Dataset	Model	Detection Performance					Segmentation Performance				
		P	R	F1	mAP_{50}	$mAP_{50–95}$	P	R	F1	mAP_{50}	$mAP_{50–95}$
Deformation rate image dataset	Yolov8	81.40	64.00	71.66	73.60	44.00	79.80	59.20	67.97	69.10	32.30
	Yolov9	94.60	87.00	90.64	93.30	72.80	93.50	82.60	87.71	90.20	57.50
	Yolov11	85.10	67.80	75.47	77.00	48.60	83.50	63.80	72.33	73.00	36.50
	Yo1ov12	85.50	67.40	75.38	77.40	50.10	85.10	65.60	74.09	75.10	38.70
	Maskrcnn-50	73.74	78.96	76.26	78.16	54.27	66.14	71.52	68.73	70.69	37.58
	Maskrcnn-101	75.46	79.67	77.51	79.24	55.31	67.46	72.53	69.90	72.20	38.39
	Maskrcnn-152	76.65	80.83	78.68	80.12	56.15	67.44	72.96	70.09	73.15	38.94
	Our model	93.95	95.93	94.93	95.00	61.86	87.06	89.98	88.50	91.25	45.18

https://doi.org/10.3390/rs18010126

Table 3. *Cont.*

Dataset	Model	Detection Performance					Segmentation Performance				
		P	**R**	**F1**	**mAP$_{50}$**	**mAP$_{50–95}$**	**P**	**R**	**F1**	**mAP$_{50}$**	**mAP$_{50–95}$**
Two-source fusion image dataset	Yolov8	84.60	66.10	74.21	76.80	46.70	85.50	64.40	73.46	75.10	39.10
	Yolov9	96.10	90.10	93.00	95.10	78.30	93.60	87.20	90.29	93.10	63.20
	Yolov11	83.10	72.10	77.21	80.50	52.70	85.40	68.10	75.78	78.50	43.30
	Yolov12	88.50	70.90	78.73	81.81	54.50	87.80	69.40	77.52	80.40	45.60
	Maskrcnn-50	81.04	84.71	82.83	84.29	59.31	76.73	80.63	78.63	80.33	45.79
	Maskrcnn-101	81.92	85.71	83.78	85.34	60.00	78.73	82.50	80.57	82.06	46.93
	Maskrcnn-152	82.90	86.39	84.60	84.99	60.60	78.24	82.22	80.18	81.19	46.65
	Our model	96.66	97.60	97.13	97.50	71.23	93.34	94.82	94.08	95.45	56.75
Multi-source fusion image dataset	Yolov8	92.80	82.30	87.24	90.60	64.00	92.80	77.30	84.34	87.70	51.80
	Yolov9	97.20	95.70	96.44	97.20	**86.80**	95.00	90.20	92.54	95.50	**70.10**
	Yolov11	89.90	79.30	84.27	87.60	60.20	89.20	76.70	82.48	85.70	49.80
	Yolov12	91.30	77.50	83.84	87.40	60.20	91.50	74.20	81.95	85.10	49.00
	Maskrcnn-50	83.08	85.67	84.35	86.10	60.10	78.63	81.54	80.06	82.10	47.20
	Maskrcnn-101	84.51	87.78	86.11	87.30	60.40	79.84	83.03	81.40	83.50	46.80
	Maskrcnn-152	83.01	86.00	84.48	86.00	60.20	78.01	81.45	79.69	82.50	47.20
	Our model	**97.79**	**98.18**	**97.98**	**97.88**	75.48	**95.70**	**96.55**	**96.12**	**96.73**	60.95

Underlined values represent the best metric for each dataset, while bold values denote the overall best performance across all datasets.

Figures 8–10 illustrate the detection results of each model on images of varying scales across the different datasets. As observed in the figure:

(1) In the deformation rate image dataset, all models showed varying levels of missed and false detections, which were particularly pronounced for small-scale images. This is primarily attributed to the relatively small spatial extent of most active landslides in these images, posing a challenge for the models to effectively extract features. The Yolov11 model exhibited the most severe detection errors, whereas MCLD R-CNN achieved an exceptionally low missed detection rate despite a few false positives. Segmentation performance across all models was generally poor, especially in large-scale images, where models struggled to outline complete landslide contours. This suggests that deformation rate features mainly assist in localization but are insufficient for determining precise coverage and contours.

(2) In both the two-source and multi-source fusion image datasets, most models exhibited reduced missed and false detections, alongside significantly enhanced segmentation performance. Notably, models trained on multi-source fusion images demonstrated superior segmentation performance compared to those trained on two-source fusion image datasets. This indicates that multi-source imagery provides richer landslide-related features, thereby enhancing the model's detection capability. However, it is notable that detection results for identical landslide instances were inconsistent across datasets. Specifically, models like Yolov8 and Yolov12 missed certain active landslides in the fusion images that they had correctly identified in other datasets. This phenomenon was primarily concentrated in small-scale images. A likely reason is that these models extract conflicting features from the fused images, rendering it difficult to distinguish whether a target is an active landslide. In summary, utilizing fusion imagery reduces detection errors, with the multi-source fusion image dataset yielding the most significant performance gains. The quantitative improvements shown in Table 3 effectively support this conclusion.

(3) Across all datasets, our proposed model consistently maintained superior detection correctness and segmentation accuracy. This visual evidence directly demonstrates the performance advantages of our model, a finding further substantiated by the quantitative metrics analyzed above.

https://doi.org/10.3390/rs18010126

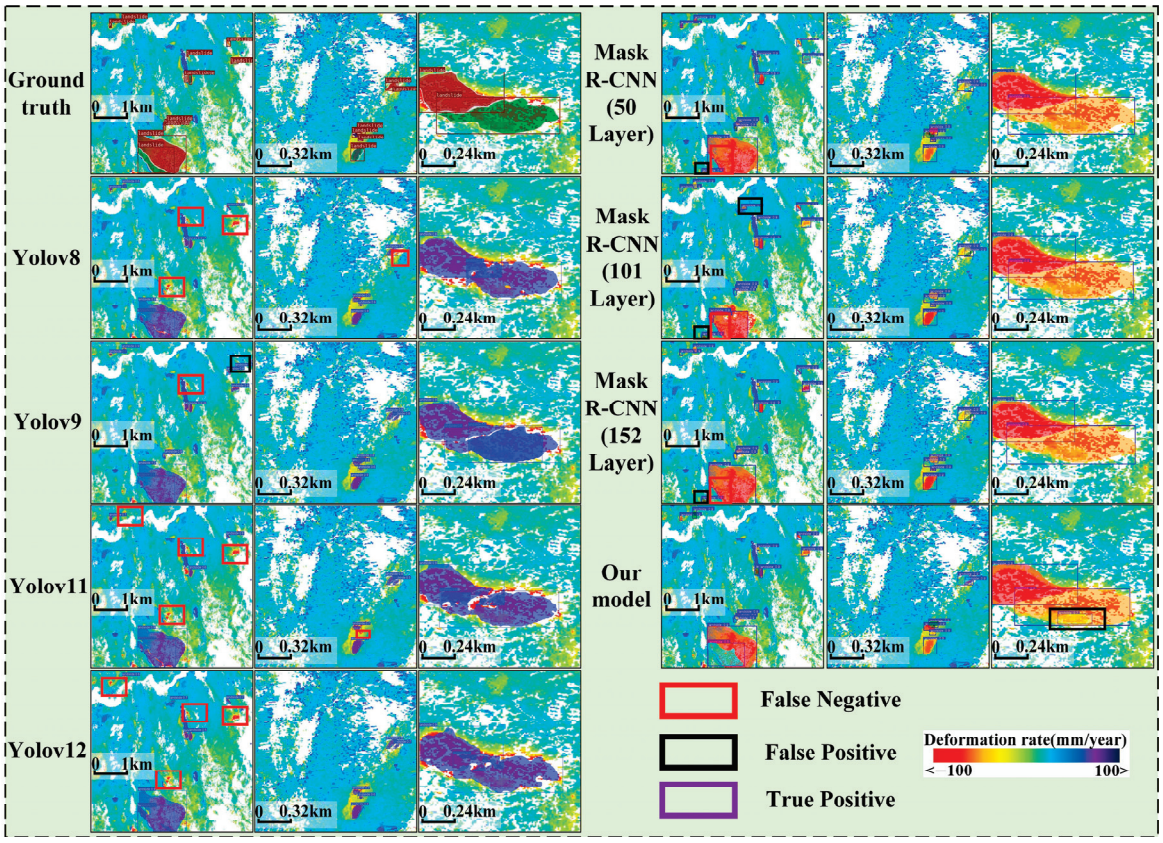


Figure 8. Detection results on the deformation rate image dataset.

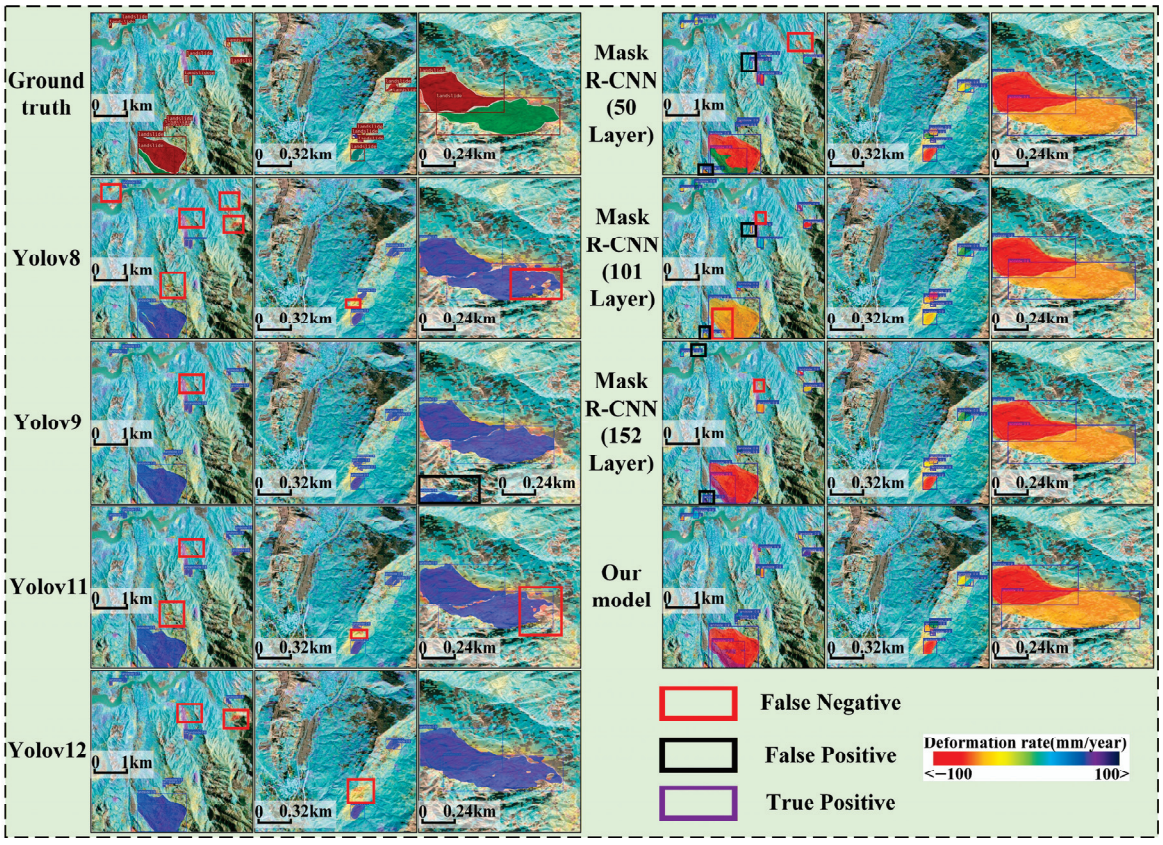


Figure 9. Detection results on the two-source fusion image dataset.

https://doi.org/10.3390/rs18010126

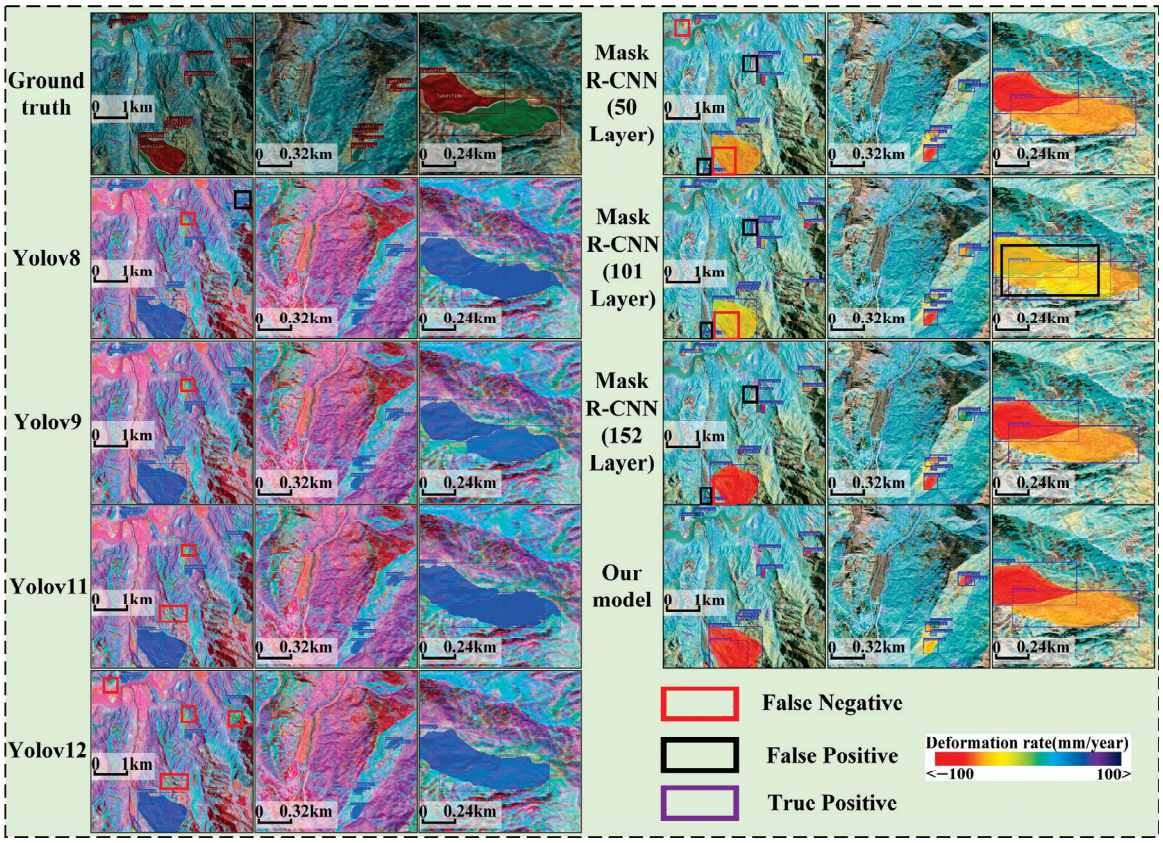


Figure 10. Detection results on the multi-source fusion image dataset. To improve visual clarity, MCLD R-CNN uses images with a transparency channel for landslide detection, but presents the prediction results on RGB-format images.

4.2.2. Model Training Curve Analysis

Figure 11 shows the mAP_{50} and Loss curves for each model across different datasets. Since the training epochs vary across models, data from the first 100 epochs are selected for comparison. As observed in the figure:

(1) Within the same model, models trained with fusion images generally exhibit better performance and faster convergence. As the variety of fused data increases, the training results progressively improve. This is particularly evident in the Yolo series models. Therefore, using multi-source fused images for training significantly enhances training efficiency, enabling the model to attain improved performance with fewer training epochs.

(2) Comparing Figure 11 with Table 3, it is clear that both MCLD R-CNN and Yolov9 models exhibit superior detection performance, but the former converges faster. Although Mask R-CNN has the fastest convergence speed, its detection performance is relatively weaker. Therefore, the model proposed in this study achieves a balance between convergence speed and detection performance. By leveraging multi-source fusion images, the training time of detection models can be reduced, thereby enhancing the efficiency of landslide detection.

https://doi.org/10.3390/rs18010126

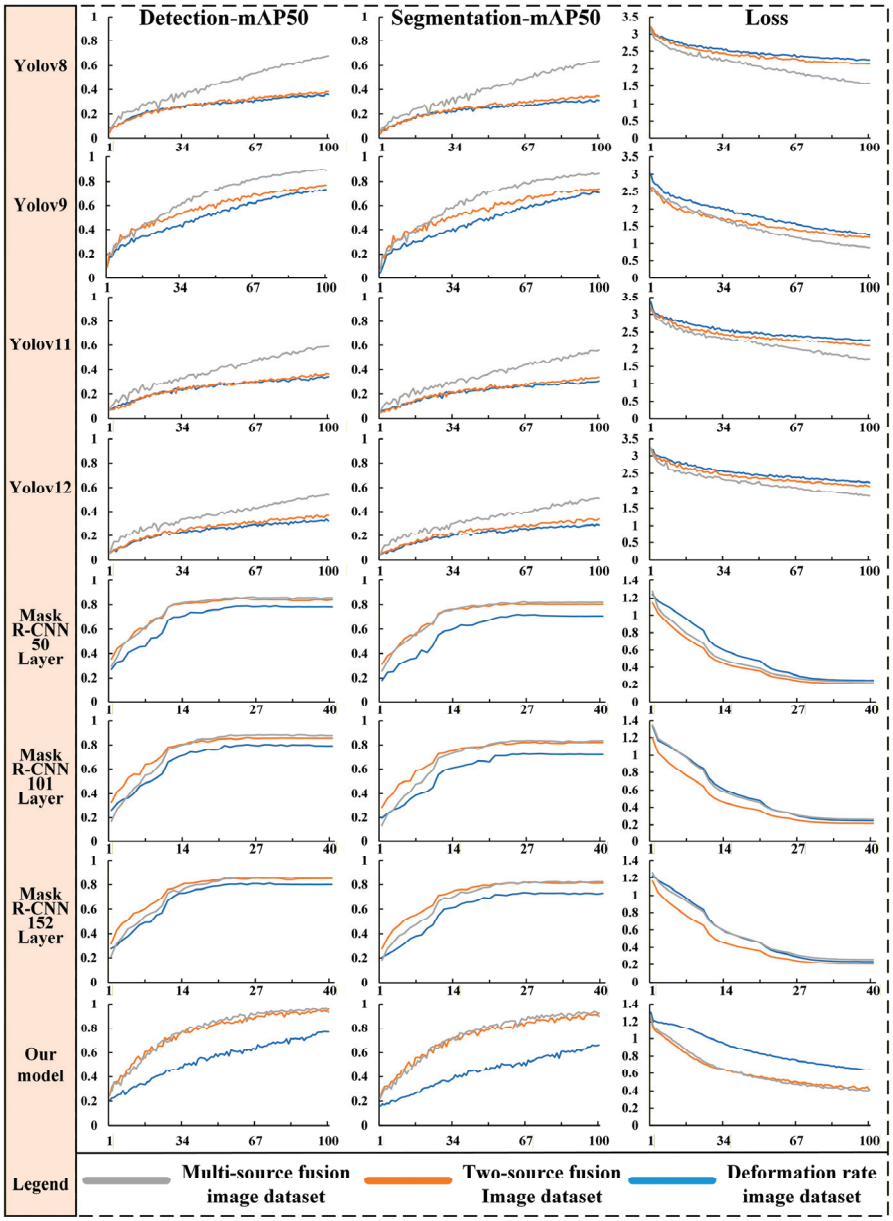


Figure 11. Training curves of each model for the first 100 epochs across three datasets.

4.3. Ablation Experiment

The Landslide Attention Module serves as a crucial component of MCLD R-CNN, enhancing the model's capability to extract and identify landslide-related features. To analyze its contribution to the landslide detection process, we trained and evaluated a model without this module. Table 4 presents the model's performance in detecting landslides using the multi-source fusion image dataset. As indicated in the table, eliminating the Landslide Attention Module resulted in a noticeable reduction in the model's detection and segmentation performance across all evaluation metrics, with the $mAP_{50–95}$ metric experiencing the largest decrease, dropping by 8.7% (detection performance) and 9.8% (segmentation performance). This finding suggests that the Landslide Attention Module is essential for enhancing the overall performance of the model. Figure 12 displays the attention heatmaps generated by the model along with the corresponding detection results. As shown in the figure:

(1) The visualization demonstrates that the attention module operates as a dual-functional filter. While it intensifies the feature response in target regions (indicated by "hot" red

https://doi.org/10.3390/rs18010126

colors), it simultaneously suppresses activation in non-landslide regions (indicated by "cool" blue colors). The Sigmoid activation function within the module acts as a gate, assigning near-zero weights to complex background features such as rivers, vegetation, and stable slopes. This effectively "mutes" the interference from these areas, preventing them from being misclassified as foreground.

(2) After removing the Landslide Attention Module, the model's attention becomes randomly dispersed across the entire image, making it difficult to focus on the target. As a result, the model struggles to detect the landslide accurately and is more easily affected by irrelevant regions.

(3) In certain cases, the model also focuses its attention on incorrect regions. For example, in the seventh sample image shown, the model concentrates on the river instead of the active landslide.

Table 4. Quantitative results of the ablation experiment.

Index	Att	Layers	Par	FLOPs	P	R	mAP_{50}	$mAP_{50–95}$	F1
Det	No	50	43.98 M	129.46 G	96.48	97.60	97.26	68.94	97.03
	Yes	50	46.51 M	201.97 G	97.79	98.18	97.88	75.48	97.98
Seg	No	50	43.98 M	129.46 G	92.35	93.82	94.52	54.99	93.07
	Yes	50	46.51 M	201.97 G	95.70	96.55	96.73	60.95	96.12

Att denotes the attention module, Par denotes the parameter count, and FLOPs corresponds to the computational complexity.

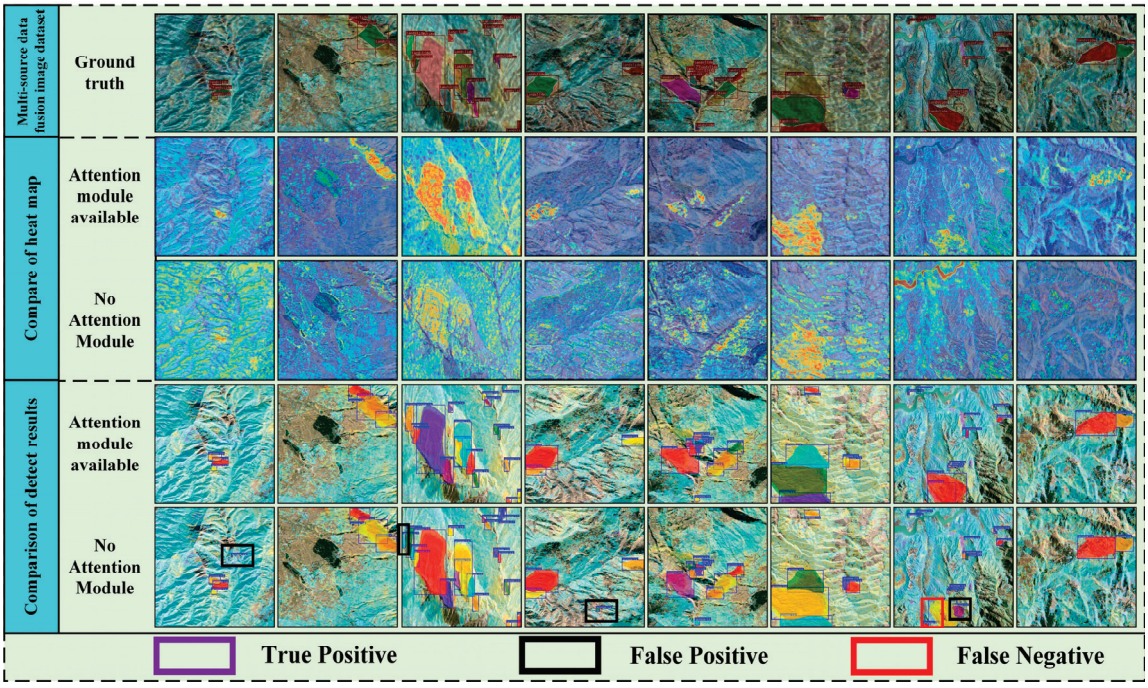


Figure 12. Heatmaps and detection results before and after introducing the landslide attention mechanism.

In summary, the model incorporating the Landslide Attention Module exhibits fewer missed and false detections, along with superior segmentation performance. This suggests that the module enables the model to concentrate more precisely on landslide areas while improving its capacity to identify and delineate landslide contours. Although the integration of the Landslide Attention Module leads to a higher computational burden and an increased number of parameters, as shown in Table 4, this trade-off is justified by its significant enhancement of model performance.

https://doi.org/10.3390/rs18010126

4.4. Evaluation of the Model's Generalization Capability

To assess the model's generalization ability, we selected the Longyangxia Reservoir area as an independent test site. This region is geographically disjoint from the training tiles. Using the established operational definition, we constructed a generalization test dataset comprising 300 image samples containing a total of 58 labeled active landslides.

Compared to the training region, the Longyangxia test site exhibits significantly more rugged terrain, featuring deep alpine canyons and steep bedrock slopes along the reservoir banks. Regarding land cover and climate, while both regions share a semi-arid climate with sparse vegetation, the test site presents more complex background interference due to the presence of the large water body (reservoir). Finally, regarding data acquisition, Sentinel-1 imagery was collected using identical acquisition parameters (IW mode, VV + VH) as the training data to strictly evaluate model transferability without sensor-related bias.

All models trained on the multi-source fusion image dataset were tested in this section, and the results are illustrated in Figure 13. As depicted in the figure:

(1) MCLD R-CNN effectively detects most active landslides, demonstrating its high sensitivity to active landslides.

(2) The Yolo series models exhibit a high number of missed detections, while Mask R-CNN exhibits poor segmentation performance. This indicates that these models struggle to adapt to unfamiliar images and can only achieve good performance on specific datasets.

(3) The quantitative results presented on the right demonstrate that variations in landslide characteristics and data distribution relative to the training datasets led to decreased performance across all models. However, MCLD R-CNN still outperformed the others overall, with scores higher than the other models for all metrics except P.

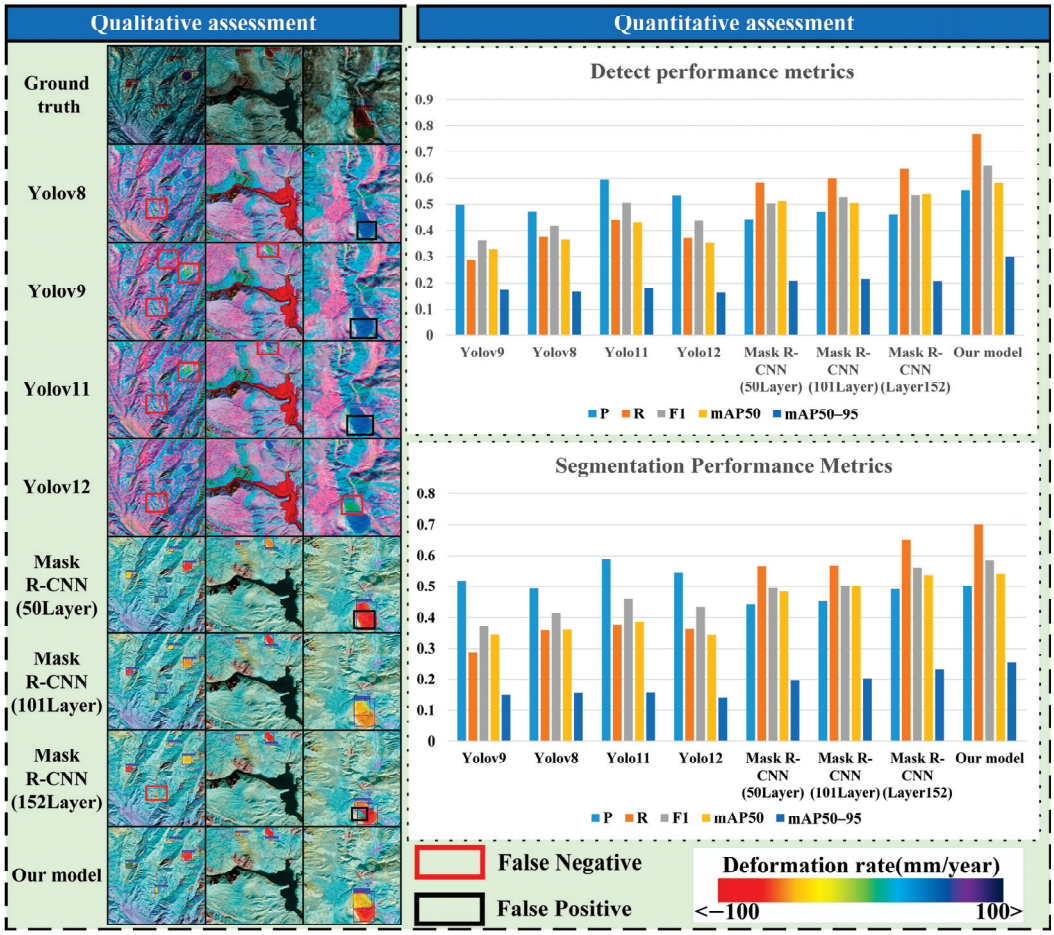


Figure 13. Model Generalization Capability Test Results.

https://doi.org/10.3390/rs18010126

In conclusion, compared to other models, MCLD R-CNN demonstrates better generalization capability. For detection tasks in specific regions, a small number of local samples can be used to fine-tune the model, enhancing its ability to detect active landslides in those particular areas.

Figure 14 presents the InSAR deformation maps and optical imagery of typical cases identified within the test region. For the successfully detected active landslides, the boundaries delineated by the proposed model (white contour lines) align well with the actual terrain and topographic features, demonstrating the robustness and reliability of the model in identifying active landslides. However, the figure also illustrates typical failure cases (missed detections) to analyze the model's limitations. As observed in the bottom row of Figure 14, these missed landslides are characterized by extremely sparse InSAR measurement points (coherent pixels) on the slope surface. The scarcity of deformation signals results in incomplete feature information, making it difficult for the model to distinguish the landslide from the background or to verify its activity status. This indicates that while the model generalizes well, its performance remains sensitive to the density and quality of the input InSAR observations.

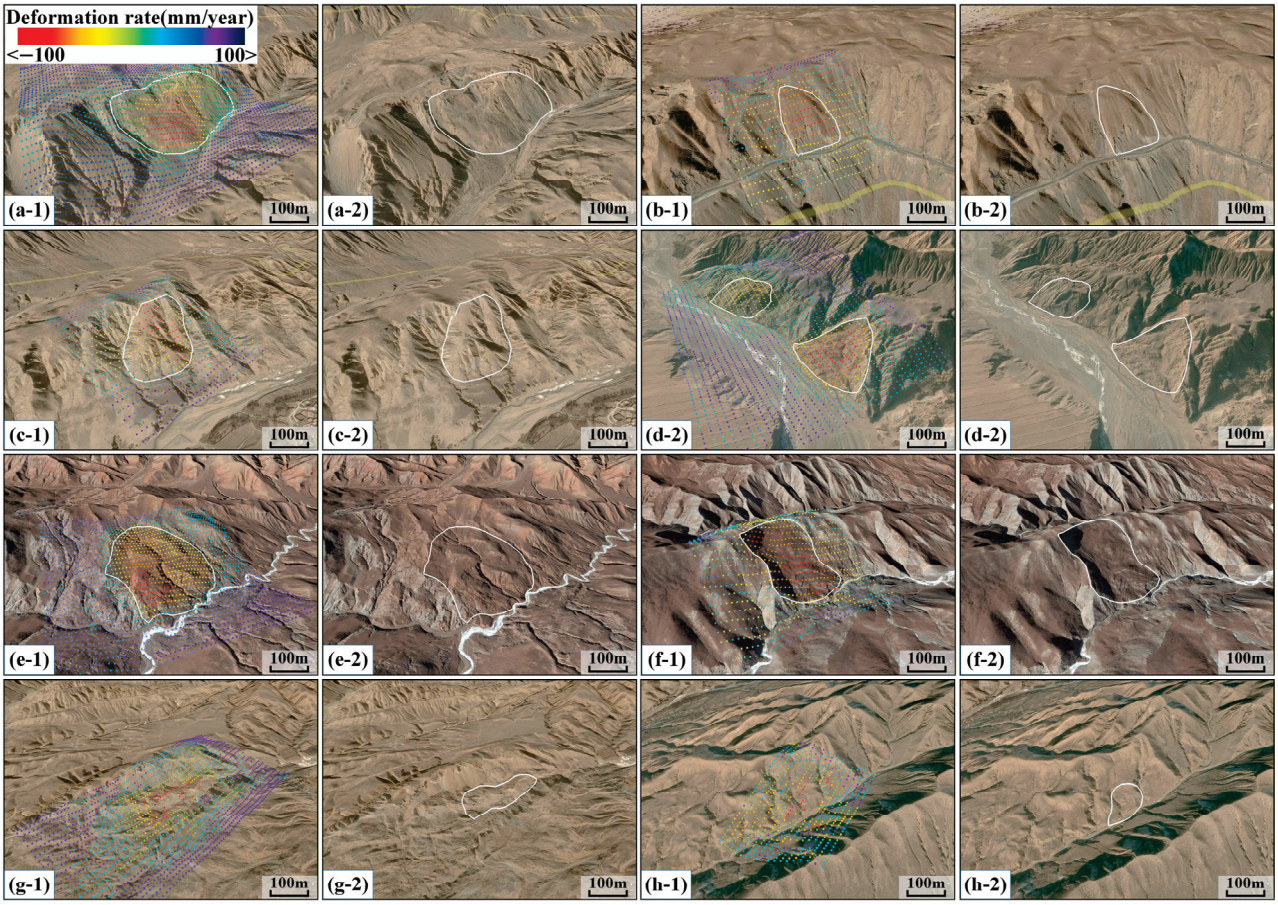


Figure 14. InSAR results (subgraphs with suffix-1) and Google Earth optical images (subgraphs with suffix-2) of representative landslide cases in the test area. (**a**–**f**) illustrate successfully detected active landslides, where white lines indicate the delineated boundaries. (**g**,**h**) represent failure cases (missed detections).

5. Discussion

5.1. Advantages and Limitations of the Dataset and Method

Multi-source fusion image enables the efficient integration of heterogeneous information within a single tensor, with distinct remarkable advantages in active landslide

https://doi.org/10.3390/rs18010126

detection tasks. Integrating multiple data modalities markedly strengthens the model's capacity to identify landslide boundaries, deformation features, and fine-scale geomorphic variations. Moreover, such fused data substantially improves model convergence speed and training stability. This allows the network to achieve superior performance with fewer training epochs, which is especially valuable under limited computational resources.

Specifically, this study fuses multi-source data, including optical imagery, DEM-derived slope, and InSAR deformation. By integrating these along the channel dimension, we adopt a single-branch network for training. This design not only maintains computational efficiency but also systematically mitigates two persistent challenges in InSAR-based detection.

Firstly, it addresses the difficulty existing methods face in distinguishing active landslides from other ground deformations, thereby reducing false alarms. By enhancing data richness through fusion, the model can better locate active landslides. This improvement is quantitatively confirmed by the significant increase in the Precision (detection performance) of MCLD R-CNN. Specifically, precision rose from 93.95% on deformation-only data to 97.79% on multi-source fusion data, indicating a substantial reduction in false positives. Simultaneously, this approach improves boundary delineation, which is often limited by blurred edges in deformation maps. By incorporating optical texture and slope information, our method enhances feature representation and sharpens boundary definition. The increase in Segmentation $mAP_{50–95}$ of the proposed model (rising from 45.18% on deformation-only data to 60.95% on multi-source fusion data) serves as robust proof of this improvement.

Furthermore, to address the challenge of class imbalance, specific strategies were employed at the data level. Rigorous data augmentation techniques, including random vertical/horizontal flipping and rotation, were applied to the training dataset. This approach effectively increased the frequency and diversity of positive samples, directly contributing to the model's notable balance between high Precision and Recall.

Nevertheless, this approach introduces greater complexity in data preprocessing, imposes higher demands on spatial co-registration accuracy, and may cause information loss during the fusion process, leading to partial feature attenuation. Future research will focus on reducing data loss, improving spatial alignment precision, and optimizing multimodal fusion strategies, with the goal of further enhancing model robustness and accuracy in complex geological environments.

5.2. Model Advantages and Limitations

Although MCLD R-CNN demonstrates superior detection and segmentation performance across three datasets as well as robust generalization capabilities, its performance on the high-precision localization metric $mAP_{50–95}$ is slightly lower than that of Yolov9. This discrepancy in localization accuracy is primarily attributed to two factors. First, differences in model architecture and regression loss play a role. Specifically, the Smooth L1 loss used by the two-stage MCLD R-CNN is less sensitive to fine-grained boundary adjustments for high-IoU samples than the dynamic label assignment and high-order IoU losses employed by Yolov9. Second, inherent limitations exist within the feature, as the bilinear interpolation in the RoI Align layer may cause slight feature smoothing at object edges. However, it is worth noting that compared to the standard Mask R-CNN baseline, the proposed modifications significantly mitigate this smoothing effect. As evidenced in Table 3, MCLD R-CNN achieves a Segmentation $mAP_{50–95}$ of 60.95%, a substantial improvement over the baseline's 47.20%, indicating that the Landslide Attention Module effectively enhances edge feature representation before pooling.

https://doi.org/10.3390/rs18010126

Furthermore, through qualitative visual assessment (as seen in Figures 8–10 and 13), we observed that while baseline models like Yolov9 tend to generate smoothed, convex masks to maximize IoU. In contrast, MCLD R-CNN produces boundaries with higher topographic conformity. It more accurately traces complex geological features. From a unified analytical perspective, MCLD R-CNN achieves a significantly higher Recall score than Yolov9, despite a slightly lower $mAP_{50–95}$. This reveals a distinct trade-off between extreme localization precision and comprehensive target capture. For active landslide detection, high recall holds greater practical value, as timely identifying potential hazards—even with slight positional deviations—is far more critical than missing high-risk targets. Future work will focus on addressing these limitations to further improve boundary optimization strategies.

6. Conclusions

This study constructed a multi-source fusion image dataset integrating optical remote sensing imagery, DEM-derived slope data, and InSAR deformation information for active landslide detection. Based on this dataset, a multi-channel instance segmentation model, termed MCLD R-CNN, was developed. The model was systematically evaluated and compared across three independent datasets. Experimental results demonstrate that training with the multi-source fusion image dataset significantly enhances both the detection performance and training convergence efficiency. These findings confirm the dataset's effectiveness and superiority in active landslide detection. Across all evaluation metrics, MCLD R-CNN achieved the best overall performance. The incorporation of the Landslide Attention Module substantially enhanced the extraction and recognition of landslide-related features, resulting in higher sensitivity and responsiveness to active landslide regions. Moreover, MCLD R-CNN exhibited strong cross-dataset generalization capability, effectively adapting to diverse data sources and geological environments. Overall, the proposed approach provides a promising framework for large-scale, high-precision, and automated detection of active landslides, offering important scientific and practical value for the monitoring of landslide hazards and early warning.

Future work will aim to enhance the diversity of the fused datasets and optimize the efficiency of data fusion. In particular, we plan to introduce a dynamic data fusion mechanism and expand the input dimensionality of the multi-channel model to integrate additional sensing modalities. These efforts are intended to improve the model's capacity for feature representation and its detection accuracy in complex and heterogeneous landslide environments.

Author Contributions: Conceptualization, J.W. and Y.R.; methodology, J.W.; software, J.W. and Y.R.; validation, J.W.; writing—original draft preparation, J.W.; formal analysis, H.F.; investigation, Y.R.; project administration, H.F.; resources, H.F. and W.T.; supervision, H.F.; data curation, Y.R. and H.F.; writing—review and editing, H.F.; visualization, J.W.; funding acquisition, H.F. and W.T. All authors have read and agreed to the published version of the manuscript.

Funding: This study was funded by the National Natural Science Foundation of China (Grant No. 42274054) and the Qinghai Institute of Technology "Kunlun Talent" Talent Introduction Research Project (No. 2023-QLGKLYCZX-25).

Data Availability Statement: Data will be made available upon request.

Acknowledgments: The authors gratefully acknowledge the European Space Agency for supplying the Sentinel-1 data.

Conflicts of Interest: The authors declare no conflicts of interest.

https://doi.org/10.3390/rs18010126

References

1. Keefer, D.K.; Larsen, M.C. Assessing landslide hazards. *Science* **2007**, *316*, 1136–1138. [CrossRef]
2. Froude, M.J.; Petley, D.N. Global fatal landslide occurrence from 2004 to 2016. *Nat. Hazards Earth Syst. Sci.* **2018**, *18*, 2161–2181. [CrossRef]
3. Lan, H.; Liu, X.; Li, L.; Li, Q.; Tian, N.; Peng, J. Remote Sensing Precursors Analysis for Giant Landslides. *Remote Sens.* **2022**, *14*, 4399. [CrossRef]
4. Xu, Q.; Zhao, B.; Dai, K.; Dong, X.; Li, W.; Zhu, X.; Yang, Y.; Xiao, X.; Wang, X.; Huang, J.; et al. Remote sensing for landslide investigations: A progress report from China. *Eng. Geol.* **2023**, *321*, 107156. [CrossRef]
5. Santangelo, M.; Cardinali, M.; Rossi, M.; Mondini, A.C.; Guzzetti, F. Remote landslide mapping using a laser rangefinder binocular and GPS. *Nat. Hazards Earth Syst. Sci.* **2010**, *10*, 2539–2546. [CrossRef]
6. Guzzetti, F.; Mondini, A.C.; Cardinali, M.; Fiorucci, F.; Santangelo, M.; Chang, K.-T. Landslide inventory maps: New tools for an old problem. *Earth Sci. Rev.* **2012**, *112*, 42–66. [CrossRef]
7. Haneberg, W.C.; Cole, W.F.; Kasali, G. High-resolution lidar-based landslide hazard mapping and modeling, UCSF Parnassus Campus, San Francisco, USA. *Bull. Eng. Geol. Environ.* **2009**, *68*, 263–276. [CrossRef]
8. Kurtz, C.; Stumpf, A.; Malet, J.-P.; Gancarski, P.; Puissant, A.; Passat, N. Hierarchical extraction of landslides from multiresolution remotely sensed optical images. *ISPRS J. Photogramm. Remote Sens.* **2014**, *87*, 122–136. [CrossRef]
9. Pesci, A.; Teza, G.; Casula, G.; Loddo, F.; De Martino, P.; Dolce, M.; Obrizzo, F.; Pingue, F. Multitemporal laser scanner-based observation of the Mt. Vesuvius crater: Characterization of overall geometry and recognition of landslide events. *ISPRS J. Photogramm. Remote Sens.* **2011**, *66*, 327–336. [CrossRef]
10. Ferretti, A.; Prati, C.; Rocca, F. Permanent scatterers in SAR interferometry. *IEEE Trans. Geosci. Remote Sens.* **2001**, *39*, 8–20. [CrossRef]
11. Massonnet, D.; Rossi, M.; Carmona, C.; Adragna, F.; Peltzer, G.; Feigl, K.; Rabaute, T. The displacement field of the Landers earthquake mapped by radar interferometry. *Nature* **1993**, *364*, 138–142. [CrossRef]
12. Zebker, H.A.; Villasenor, J. Decorrelation in interferometric radar echoes. *IEEE Trans. Geosci. Remote Sens.* **1992**, *30*, 950–959. [CrossRef]
13. Zhuang, H.; Tang, Z.; Du, S.; Wang, P.; Fan, H.; Hao, M.; Tan, Z. Shadow-robust unsupervised flood mapping via GMM-enhanced generalized dual-polarization flood index and topography features. *Int. J. Appl. Earth Obs. Geoinf.* **2025**, *143*, 104787. [CrossRef]
14. Zhuang, H.; Wang, P.; Hao, M.; Fan, H.; Tan, Z. Flood inundation mapping in SAR images based on nonlocal polarization combination features. *J. Hydrol.* **2025**, *646*, 132326. [CrossRef]
15. Pedretti, L.; Bordoni, M.; Vivaldi, V.; Figini, S.; Parnigoni, M.; Grossi, A.; Lanteri, L.; Tararbra, M.; Negro, N.; Meisina, C. InterpolatiON of InSAR Time series for the dEtection of ground deforMatiOn eVEnts (ONtheMOVE): Application to slow-moving landslides. *Landslides* **2023**, *20*, 1797–1813. [CrossRef]
16. Rosen, P.A.; Hensley, S.; Joughin, I.R.; Li, F.K.; Madsen, S.N.; Rodriguez, E.; Goldstein, R.M. Synthetic aperture radar interferometry. *Proc. IEEE* **2000**, *88*, 333–382. [CrossRef]
17. Zhou, C.; Cao, Y.; Gan, L.; Wang, Y.; Motagh, M.; Roessner, S.; Hu, X.; Yin, K. A novel framework for landslide displacement prediction using MT-InSAR and machine learning techniques. *Eng. Geol.* **2024**, *334*, 107497. [CrossRef]
18. Dille, A.; Kervyn, F.; Handwerger, A.L.; d'Oreye, N.; Derauw, D.; Bibentyo, T.M.; Samsonov, S.; Malet, J.-P.; Kervyn, M.; Dewitte, O. When image correlation is needed: Unravelling the complex dynamics of a slow-moving landslide in the tropics with dense radar and optical time series. *Remote Sens. Environ.* **2021**, *258*, 112402. [CrossRef]
19. Kang, Y.; Lu, Z.; Zhao, C.; Qu, W. Inferring slip-surface geometry and volume of creeping landslides based on InSAR: A case study in Jinsha River basin. *Remote Sens. Environ.* **2023**, *294*, 113620. [CrossRef]
20. Lauknes, T.R.; Piyush Shanker, A.; Dehls, J.F.; Zebker, H.A.; Henderson, I.H.C.; Larsen, Y. Detailed rockslide mapping in northern Norway with small baseline and persistent scatterer interferometric SAR time series methods. *Remote Sens. Environ.* **2010**, *114*, 2097–2109. [CrossRef]
21. Wang, W.; Motagh, M.; Mirzaee, S.; Li, T.; Zhou, C.; Tang, H.; Roessner, S. The 21 July 2020 Shaziba landslide in China: Results from multi-source satellite remote sensing. *Remote Sens. Environ.* **2023**, *295*, 113669. [CrossRef]
22. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
23. Ma, Z.; Mei, G. Deep learning for geological hazards analysis: Data, models, applications, and opportunities. *Earth Sci. Rev.* **2021**, *223*, 103858. [CrossRef]
24. Wei, R.; Li, Y.; Li, Y.; Zhang, B.; Wang, J.; Wu, C.; Yao, S.; Ye, C. A universal adapter in segmentation models for transferable landslide mapping. *ISPRS J. Photogramm. Remote Sens.* **2024**, *218*, 446–465. [CrossRef]
25. Lu, W.; Hu, Y.; Zhang, Z.; Cao, W. A dual-encoder U-Net for landslide detection using Sentinel-2 and DEM data. *Landslides* **2023**, *20*, 1975–1987. [CrossRef]

https://doi.org/10.3390/rs18010126

26. Lu, Z.; Peng, Y.; Li, W.; Yu, J.; Ge, D.; Han, L.; Xiang, W. An Iterative Classification and Semantic Segmentation Network for Old Landslide Detection Using High-Resolution Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 4408813. [CrossRef]
27. Gu, W.; Bai, S.; Kong, L. A review on 2D instance segmentation based on deep neural networks. *Image Vis. Comput.* **2022**, *120*, 104401. [CrossRef]
28. Lv, P.; Ma, L.; Li, Q.; Du, F. ShapeFormer: A Shape-Enhanced Vision Transformer Model for Optical Remote Sensing Image Landslide Detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 2681–2689. [CrossRef]
29. Ullo, S.L.; Mohan, A.; Sebastianelli, A.; Ahamed, S.E.; Kumar, B.; Dwivedi, R.; Sinha, G.R. A New Mask R-CNN-Based Method for Improved Landslide Detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 3799–3810. [CrossRef]
30. Cai, J.; Zhang, L.; Dong, J.; Guo, J.; Wang, Y.; Liao, M. Automatic identification of active landslides over wide areas from time-series InSAR measurements using Faster RCNN. *Int. J. Appl. Earth Obs. Geoinf.* **2023**, *124*, 103516. [CrossRef]
31. Guo, H.; Yi, B.; Yao, Q.; Gao, P.; Li, H.; Sun, J.; Zhong, C. Identification of Landslides in Mountainous Area with the Combination of SBAS-InSAR and Yolo Model. *Sensors* **2022**, *22*, 6235. [CrossRef] [PubMed]
32. Zhang, T.; Zhang, W.; Cao, D.; Yi, Y.; Wu, X. A New Deep Learning Neural Network Model for the Identification of InSAR Anomalous Deformation Areas. *Remote Sens.* **2022**, *14*, 2690. [CrossRef]
33. Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 833–851.
34. Chen, X.; Yao, X.; Zhou, Z.; Liu, Y.; Yao, C.; Ren, K. DRs-UNet: A Deep Semantic Segmentation Network for the Recognition of Active Landslides from InSAR Imagery in the Three Rivers Region of the Qinghai–Tibet Plateau. *Remote Sens.* **2022**, *14*, 1848. [CrossRef]
35. Chen, X.; Zhao, C.; Liu, X.; Zhang, S.; Xi, J.; Khan, B.A. An Embedding Swin Transformer Model for Automatic Slow-Moving Landslide Detection Based on InSAR Products. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5223915. [CrossRef]
36. Chen, H.; He, Y.; Zhang, L.; Yang, W.; Liu, Y.; Gao, B.; Zhang, Q.; Lu, J. A Multi-Input Channel U-Net Landslide Detection Method Fusing SAR Multisource Remote Sensing Data. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *17*, 1215–1232. [CrossRef]
37. Li, X.; Guo, X.; Li, W. Mechanism of Giant Landslides from Longyangxia Valley to Liujiaxia Valley along Upper Yellow River. *J. Eng. Geol.* **2011**, *19*, 516–529. [CrossRef]
38. Du, J.; Li, Z.; Song, C.; Zhu, W.; Ji, Y.; Zhang, C.; Chen, B.; Su, S. InSAR-Based Active Landslide Detection and Characterization Along the Upper Reaches of the Yellow River. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 3819–3830. [CrossRef]
39. Peng, J.; Lan, H.; Qian, H.; Wang, W.; Li, R.; Li, Z.; Zhuang, J.; Liu, X.; Liu, S. Scientific Research Framework of Livable Yellow River. *J. Eng. Geol.* **2020**, *28*, 189–201. [CrossRef]
40. Yin, Z.; Cheng, G.; Hu, G.; Wei, G.; Wang, Y. Preliminary Study on Characteristic and Mechanism of Super-Large Landslides in Upper Yellow River Since Late-Pleistocene. *J. Eng. Geol.* **2010**, *18*, 41. [CrossRef]
41. Yin, Z.; Qin, X.; Zhao, W. Characteristics of Landslides from Sigou Gorge to Lagan Gorge in the Upper Reaches of Yellow River. In Proceedings of the Third World Landslide Forum, Beijing, China, 2–6 June 2014; pp. 397–406.
42. Berardino, P.; Fornaro, G.; Lanari, R.; Sansosti, E. A new algorithm for surface deformation monitoring based on small baseline differential SAR interferograms. *IEEE Trans. Geosci. Remote Sens.* **2002**, *40*, 2375–2383. [CrossRef]
43. Wegmller, U.; Strozzi, T.; Wiesmann, A. GAMMA SAR and interferometric processing software. In Proceedings of the ERS-Envisat Symposium, Gothenburg, Sweden, 16–20 October 2000.
44. Xue, F.; Lv, X.; Dou, F.; Yun, Y. A Review of Time-Series Interferometric SAR Techniques: A Tutorial for Surface Deformation Analysis. *IEEE Geosci. Remote Sens. Mag.* **2020**, *8*, 22–42. [CrossRef]
45. Hooper, A. A multi-temporal InSAR method incorporating both persistent scatterer and small baseline approaches. *Geophys. Res. Lett.* **2008**, *35*, L16302. [CrossRef]
46. Li, Y.; Zhang, Y.; Su, X.; Zhao, F.; Liang, Y.; Meng, X.; Jia, J. Early identification and characteristics of potential landslides in the Bailong River Basin using InSAR technique. *Natl. Remote Sens. Bull.* **2021**, *25*, 677–690. [CrossRef]
47. He, Y.; Wang, W.; Zhang, L.; Chen, Y.; Chen, Y.; Chen, B.; He, X.; Zhao, Z. An identification method of potential landslide zones using InSAR data and landslide susceptibility. *Geomat. Nat. Hazards Risk* **2023**, *14*, 2185120. [CrossRef]
48. Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
49. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2980–2988.
50. Brabandere, B.D.; Neven, D.; Gool, L.V. Semantic Instance Segmentation with a Discriminative Loss Function. *arXiv* **2017**. [CrossRef]
51. Hariharan, B.; Arbeláez, P.; Girshick, R.; Malik, J. Simultaneous Detection and Segmentation. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 297–312.

https://doi.org/10.3390/rs18010126

52. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
53. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the 15th European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
54. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 936–944.
55. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]
56. Varghese, R.; M., S. YOLOv8: A Novel Object Detection Algorithm with Enhanced Performance and Robustness. In Proceedings of the 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS), Chennai, India, 18–19 April 2024; pp. 1–6.
57. Wang, C.-Y.; Yeh, I.-H.; Mark Liao, H.-Y. YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information. In Proceedings of the European Conference on Computer Vision, Milan, Italy, 29 September–4 October 2024; pp. 1–21.
58. Khanam, R.; Hussain, M. YOLOv11: An Overview of the Key Architectural Enhancements. *arXiv* **2024**. [CrossRef]
59. Tian, Y.; Ye, Q.; Doermann, D. YOLOv12: Attention-Centric Real-Time Object Detectors. *arXiv* **2025**. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

https://doi.org/10.3390/rs18010126

remote sensing

MDPI

Article

A Hybrid Physics–Machine Learning Framework for Landslide Susceptibility Assessment with an Improved Non–Landslide Sampling Strategy

Dalei Peng [1], Maoyuan Chen [1,*], Yeping Zhou [1], Pinliang Li [1], Shihao Xiao [2], Yuyang Shen [1], Boren Tan [1], Linghao Kong [1] and Qiang Xu [1]

[1] State Key Laboratory of Geohazard Prevention and Geoenvironment Protection, Chengdu University of Technology, Chengdu 610059, China; pengdalei@cdut.edu.cn (D.P.); zhouyeping@stu.cdut.edu.cn (Y.Z.); lpl24@stu.cdut.edu.cn (P.L.); shenyuyang@stu.cdut.edu.cn (Y.S.); boren@stu.cdut.edu.cn (B.T.); konglinghao@stu.cdut.edu.cn (L.K.); xq@cdut.edu.cn (Q.X.)

[2] Department of Civil and Environmental Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, China; sxiaoai@connect.ust.hk

* Correspondence: cmy@stu.cdut.edu.cn

Highlights

What are the main findings?

- An improved non–landslide sampling strategy increases landslide susceptibility prediction accuracy by 16.46%, with the MLP model outperforming RF, SVM, and XGBoost.
- The proposed hybrid physics–machine learning framework integrates key factors like rainfall, groundwater, and human activity, aligning with SHAP analysis findings.

What are the implications of the main findings?

- The improved non–landslide sampling strategy coupled with MLP offers a superior alternative to the traditional buffering method for evaluating landslide susceptibility.
- The proposed framework successfully incorporates physical mechanisms of clustered landslides into regional assessments, as demonstrated by the 2024 Typhoon Gaemi event.

Abstract

Rainfall–triggered clustered landslides pose severe risks to communities and infrastructure in mountainous regions. High–precision susceptibility assessment is essential for early warning and hazard mitigation. The traditional buffering method neglects physical slope stability mechanisms, leading to the misclassification of potentially unstable areas. To improve susceptibility model accuracy, we propose an improved non–landslide sampling strategy that integrates the physical–model TRIGRS (Transient Rainfall Infiltration and Grid–based Regional Slope–Stability Model) with 50 m buffering constraints. A hybrid physics–machine learning framework is used to evaluate the performance of landslide susceptibility assessment across four machine learning models, such as Multi–Layer Perceptron (MLP), Random Forest (RF), Support Vector Machine (SVM), and Extreme Gradient Boosting (XGBoost). Among the four models, the TRIGRS model integrated with MLP achieves the highest accuracy in susceptibility mapping. The improved non–landslide sampling strategy increased average Area Under the Curve (AUC) by 16.46% in random cross–validation and improved spatial generalization capability by 29% in spatial cross–validation, demonstrating its robustness in unseen areas. SHAP factor analysis further confirms rainfall, groundwater table, and human activity as the primary influencing factors, which

aligns with physical mechanisms and improves model interpretability. Therefore, the proposed non–landslide sampling strategy coupled with the TRIGRS and MLP models outperforms traditional buffering method in evaluating regional landslide susceptibility, providing a more physically basis for geohazard risk assessment.

Keywords: rainfall–induced landslides; landslide susceptibility; physics–based model; non–landslide sample; sampling strategy

1. Introduction

The increasing frequency and severity of typhoons and extreme precipitation in recent years have led to a notable escalation in rainfall–triggered landslides and their associated impacts [1]. The escalating impacts of global climate change have further intensified the frequency and severity of these extreme weather events [2,3]. Consequently, areas surrounding cities, major infrastructure developments, and world heritage sites now face heightened landslide risk [4,5]. To address this challenge, landslide susceptibility assessment serves as a vital tool for estimating the spatial likelihood of landslide occurrence, thereby offering a scientific basis for risk evaluation and hazard mitigation [6–8].

The accuracy of landslide susceptibility assessments is strongly dependent on multiple factors, including the quality and resolution of input data, the choice of predictive model, and the methodological framework employed. The quality of training samples is a crucial element among them [9–11]. Previous research has made significant progress in identifying and sampling landslides, while the importance of non–landslide samples has been largely overlooked [12,13]. Non–landslide samples are crucial for delineating the decision boundaries between stable and unstable conditions in susceptibility modeling [14]. The inclusion of non–landslide samples through a random selection process has the potential to introduce incorrect labels, which can result in misclassifying potentially unstable areas as stable, despite their unfavorable geological conditions [15]. This systematic bias reduces the model capacity to generalize and weakens the robustness of predictions [16]. Therefore, achieving a scientifically sound and representative selection of non–landslide samples has emerged as a critical challenge in enhancing the accuracy of landslide susceptibility models [17].

Common non–landslide sampling approaches include random selection, susceptibility–based sampling, and buffer–zone exclusion [18]. Random sampling fails to account for spatial patterns, potentially mislabeling non–landslide samples in highly susceptible areas [19]. Susceptibility assessment sampling creates potential landslide training samples by preliminary model [20]. However, this sampling approach does not account for the movement scope and affected areas of slope failure [21,22]. Buffer zone sampling is commonly used due to its ability to partially reflect spatial impact of slope failure despite its simplicity [23]. However, the precision of this method relies on indirect indicators and the subjectively defined buffer radii [24]. These non–landslide sampling strategies have inherent limitations without considering the underlying physical mechanisms of slope failure, which may inadvertently contain unstable slope sample, increase model uncertainty and reduce its performance [25]. In this study, such a traditional buffering method is explicitly retained as a benchmark to rigorously evaluate the performance improvements offered by the improved non–landslide sampling strategy.

This study aims to: (1) develop an improved non–landslide sampling strategy by coupling a physics–based model (TRIGRS) with four machine learning algorithms; and (2) evaluate the resulting gains in landslide susceptibility model accuracy relative to traditional

https://doi.org/10.3390/rs18030408

buffering method. Using the landslide event triggered by the 2024 Typhoon Gaemi in Zixing County as a case study, we propose an improved strategy that integrates the TRIGRS model with 50 m buffering constraints. This method directly filters stable areas from a mechanistic perspective by simulating rainfall infiltration and slope stability, while also incorporating spatial exclusion to minimize interference. The findings provide a new solution to the challenge of non–landslide sample selection. High–precision susceptibility maps can directly inform dynamic regional landslide risk assessments, thereby strengthening evidence–based disaster prevention and mitigation efforts.

2. Materials

2.1. Study Area

The study area lies in southeastern Hunan Province, at the convergence of the western Luoxiao Mountains and the northern Nanling Mountains. The topography generally slopes downward from the southeast to the northwest, with elevations varying between 268 m and 2042 m (Figure 1a,b). The southeastern area is characterized by steep mountainous topography, while the northwestern region is relatively flat. The local geology is predominantly composed of Sinian, Cambrian, Devonian, and Caledonian rocks. Figure 1c depicts the dominant lithology (granite (Gr), metasandstone (Mss), siliceous rock (Crt), limestone (Ls), and quartz sandstone (Qss)), along with major faults, roads, and rivers. The hydrological system is part of the Xiangjiang River Basin. The Dongjiang Reservoir serves as a major water management hub, with a total storage capacity of 8.12×10^9 m^3 and a surface area of approximately 160 km^2 [26]. The study area experiences a humid subtropical monsoon climate, receiving an average annual precipitation of 1487.6 mm, most of which occurs between March and August. A notable event occurred from 26 to 28 July 2024, when Typhoon Gaemi passed through Xingning Town, bringing over 500 mm of rainfall and triggering numerous landslides (Figure 1d).

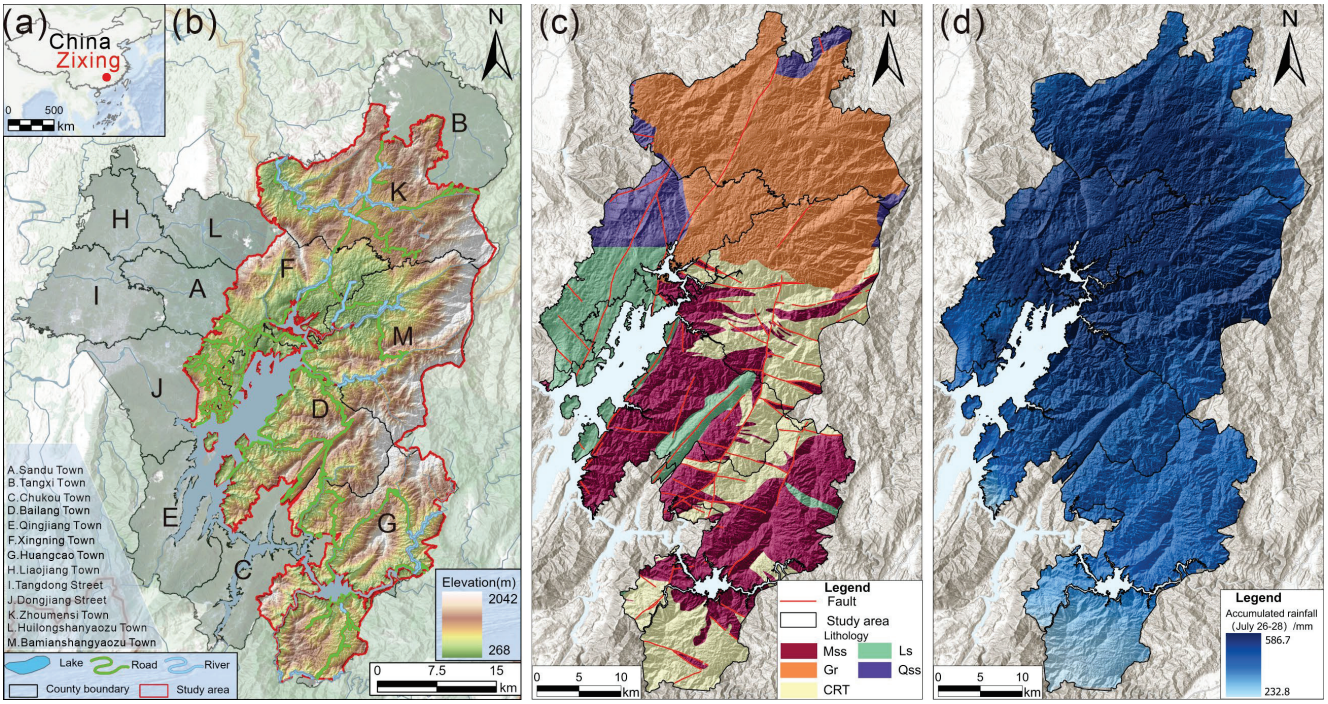


Figure 1. Location, topography, lithology, and accumulated rainfall during July 26–28 in the study area. (**a**) Location of Zixing County, Hunan Province, China; (**b**) Topography and main towns in Zixing County; (**c**) Lithology and main faults; (**d**) Accumulated rainfall (26–28 July 2024) triggered by the 2024 Typhoon Gaemi.

https://doi.org/10.3390/rs18030408

2.2. Data Source

Multi–source geospatial data sources, acquisition periods, and spatial resolutions for all geological environment and human activity used in this study are summarized in Table 1. The Digital Elevation Model (DEM) (spatial resolution 5 m) and pre–event satellite imagery (spatial resolution 2 m) were acquired on 1 May 2023, while post–event imagery (spatial resolution 0.7 m) was obtained on 24 October 2024. For vegetation analysis, the normalized difference vegetation index (NDVI) between May 2023 and October 2024 was derived from the Sentinel–2 multispectral image data with a spatial resolution of 10 m and a temporal resolution of 3–5 d [27]. Rainfall data was derived from radar–based precipitation provided by Caiyun Technology Company (Chengdu, China), with a spatial resolution of 1 km and a temporal resolution of 1 h. Monthly groundwater table data (spatial resolution 1 km) and a geological map (scale 1:200,000) were obtained from the National Geological Data Center [28]. The soil thickness was calculated using the linear slope–dependent model (Equation (1)) proposed by Saulnier et al. (1997) [29], which assumes that soil thickness was inversely proportional to the slope gradient:

$$d_i = d_{\max}\left\{1 - \left[\frac{\tan\beta_i - \tan\beta_{\min}}{\tan\beta_{\max} - \tan\beta_{\min}}\left(1 - \frac{d_{\min}}{d_{\max}}\right)\right]\right\} \quad (1)$$

where d_i represents the soil thickness at grid cell i; $d_{\max}$ and $d_{\min}$ are the maximum and minimum soil thickness in the study area, determined as 5 m and 0.8 m based on field surveys; β_i is the slope gradient at grid cell i; $\beta_{\max}$ and $\beta_{\min}$ denote the maximum and minimum slope angles in the study area, respectively.

Table 1. Data sources and spatial resolutions.

Type	Indicators	Spatial Resolution	Source
Remote sensing imagery	Pre–event images Post–event images	2.0 m 0.7 m	ZY–3 satellite–FWD Jilin–1 satellite–PMS
Topography	Elevation (m)	5 m	ZY–3 satellite–FWD
Geological environment	Normalized difference vegetation index	10 m	Sentinel–2
	Soil thickness (m)	0.5 m	Calculated based on empirical Equation (1)
	Lithology	1:200,000	National Geological Data Center (https://www.resdc.cn/)
	Accumulated rainfall (mm)	1 km	Radar–based precipitation data produced by Caiyun Technology Company
	Groundwater table (m)	1 km	National Geological Data Center (https://www.resdc.cn/)
	Distance to fault (m)	5 m	Derived from Euclidean distance to fault lines referring to https://www.resdc.cn/
	Distance to river (m)	5 m	Derived from Euclidean distance to river lines referring to https://www.resdc.cn/
Human activity	Distance to road (m)	5 m	Derived from Euclidean distance to road lines referring to https://www.resdc.cn/

3. Methods

3.1. Modelling Procedure

The methodological framework of this study is illustrated in Figure 2. The workflow was developed to enhance the accuracy of landslide susceptibility assessment and was structured into three primary stages: (1) data acquisition and processing, (2) machine learning model development, and (3) landslide probability prediction and mapping.

https://doi.org/10.3390/rs18030408

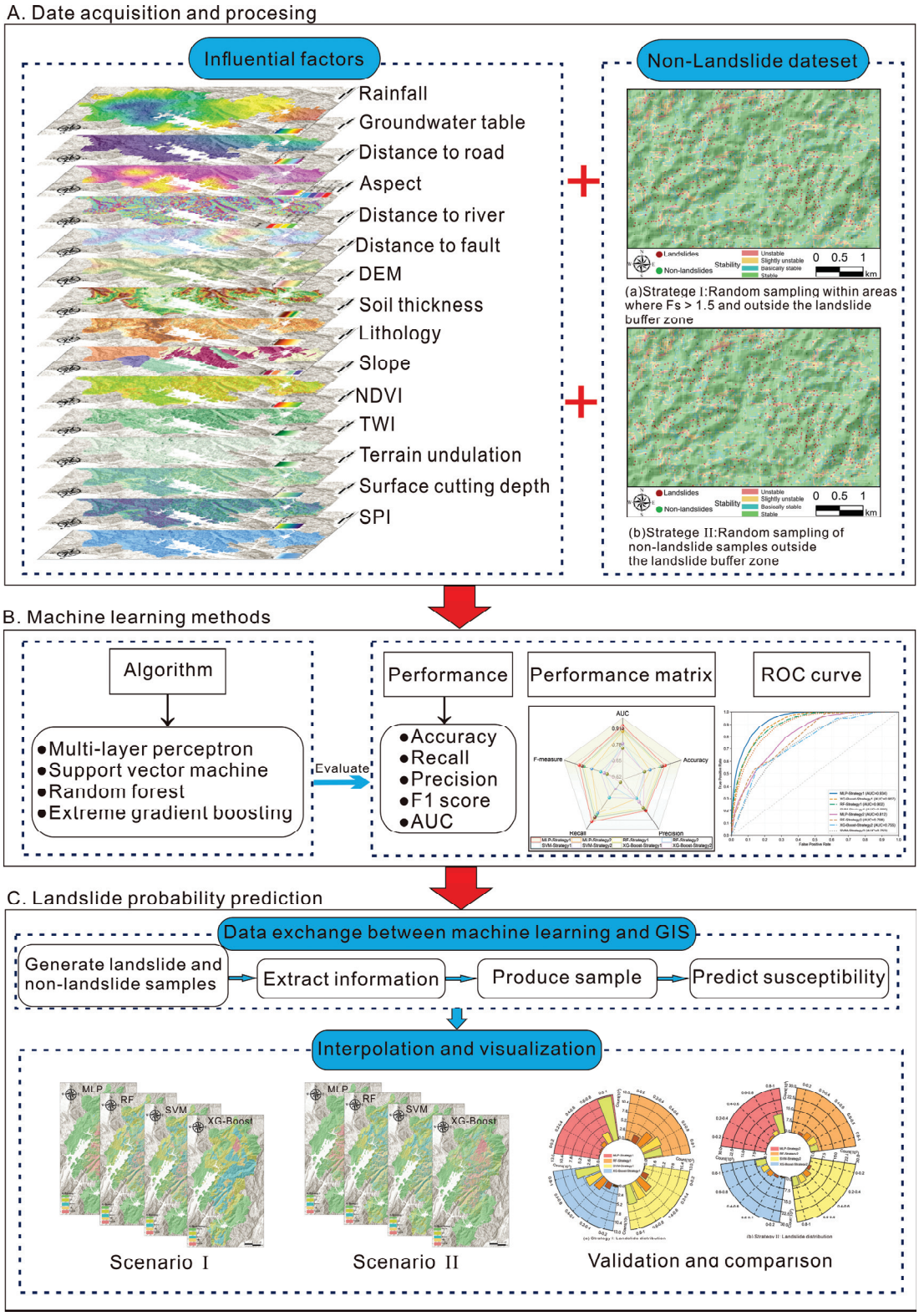


Figure 2. Methodological flowchart of this study.

Initially, a comprehensive spatial database was established by integrating the landslide inventory with 15 potential influencing factors. To ensure scale consistency across heterogeneous datasets, all input layers were resampled to a uniform spatial resolution of 5 m × 5 m. This resolution was selected to align with the high–precision DEM (5 m) and high–resolution satellite imagery (0.7 m) used in the study, which is essential for accurately capturing topographic controls on shallow landslides. Specifically, continuous variables

https://doi.org/10.3390/rs18030408

(e.g., rainfall) were processed using bilinear interpolation to preserve spatial continuity, while categorical variables (e.g., lithology) used nearest neighbor assignment.

In this study, regular grid cells were adopted as the basic mapping units. The study area was discretized into grid cells with a spatial resolution of 5 m × 5 m, consistent with the DEM data. This grid–based approach was selected to preserve the fine–scale spatial heterogeneity captured by the high–resolution data. Following grid cell discretization, training datasets were constructed by combining positive samples (17,529 landslides extracted from landslide crown areas) with an equal number of negative samples (17,529 non–landslides) selected via two distinct strategies, maintaining a balanced 1:1 ratio. These datasets were employed to train four machine learning models: MLP, SVM, RF, and XGBoost. Model performance was assessed using both random and spatial cross–validation.

3.2. Landslide Influencing Factors

The complex and diverse development and mechanism of landslides over a large area are mainly influenced by five key factors: topography, hydrology, land cover, geology, and other factors such as human activity. A total of 15 potential influencing factors in the landslide susceptibility model have been identified (Figure 3), including elevation, slope, aspect, stream power index, topographic wetness index, terrain undulation, surface cutting depth, NDVI, soil thickness, lithology, rainfall, groundwater table, distance to fault, distance to river, and distance to road.

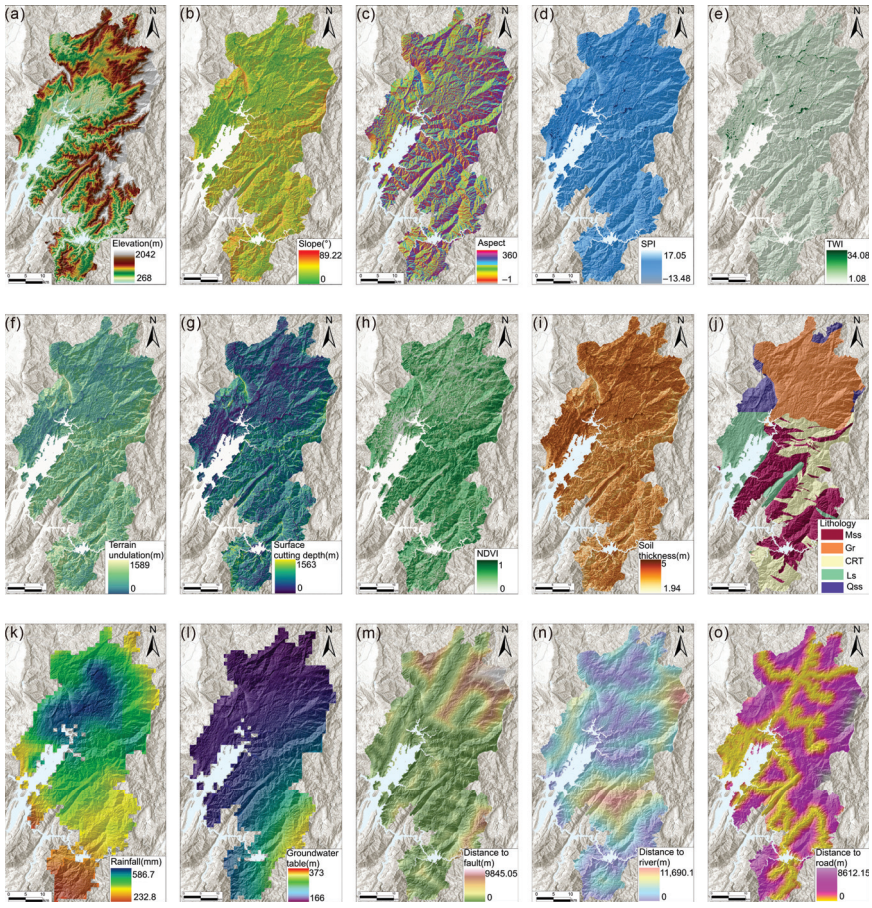


Figure 3. Spatial distribution of the fifteen landslide–influencing factors. (**a**) Elevation; (**b**) Slope; (**c**) Aspect; (**d**) Stream power index; (**e**) Topographic wetness index; (**f**) Terrain undulation; (**g**) Surface cutting depth; (**h**) Normalized Difference Vegetation Index (NDVI); (**i**) Soil thickness; (**j**) Lithology; (**k**) Accumulated rainfall during July 26–28; (**l**) Groundwater table; (**m**) Distance to fault; (**n**) Distance to river; (**o**) Distance to road.

https://doi.org/10.3390/rs18030408

To minimize information redundancy and ensure model robustness, a two–stage screening strategy was employed. Firstly, a Pearson correlation analysis was conducted, indicating that pairwise correlations among most factors were low (<0.5) (Figure 4). However, to detect potential multivariate collinearity that pairwise comparisons might miss, a variance inflation factor (VIF) analysis was performed. As presented in Table 2, the VIF values for all 15 factors range from 1.02 to 4.97, falling well below the strict threshold of 5. These results quantitatively confirm that the selected factors are statistically independent and suitable for multivariate modeling.

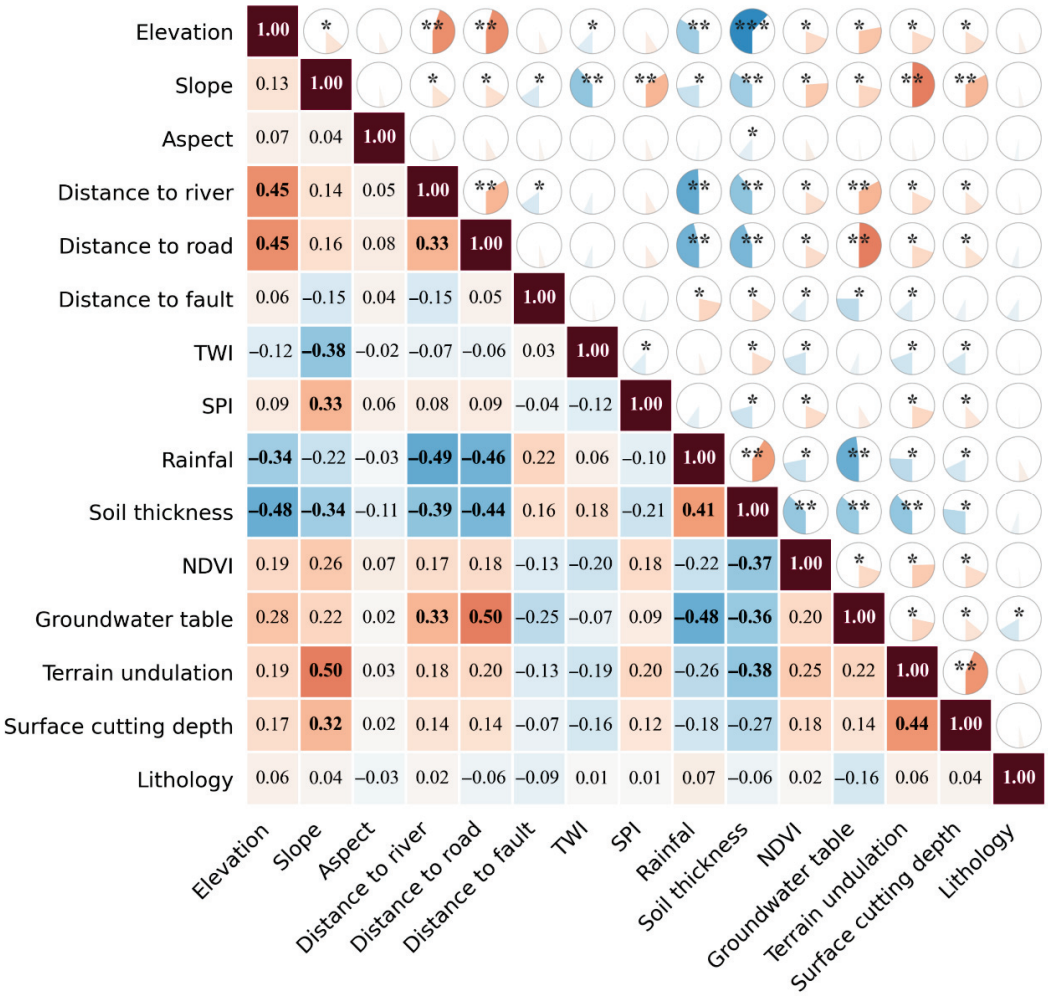


Figure 4. Heatmap of Pearson correlation coefficients among the influencing factors. The color scale indicates the correlation coefficient values, with red representing positive correlation and blue representing negative correlation. Asterisks indicate statistical significance: * $p < 0.05$, ** $p < 0.01$, and *** $p < 0.001$.

Table 2. Multicollinearity analysis of the conditioning factors.

Factor Category	Conditioning Factor	Variance Inflation Factor	Result
Topography	Terrain undulation	4.97	No multicollinearity
	Surface cutting depth	4.32	No multicollinearity
	Elevation	3.21	No multicollinearity
	Slope	2.02	No multicollinearity
	Aspect	1.02	No multicollinearity
	TWI	1.22	No multicollinearity
	SPI	1.15	No multicollinearity

https://doi.org/10.3390/rs18030408

Table 2. *Cont.*

Factor Category	Conditioning Factor	Variance Inflation Factor	Result
Hydrology	Distance to river	1.84	No multicollinearity
	Rainfall	2.01	No multicollinearity
	Groundwater table	1.88	No multicollinearity
Geology	Lithology	1.08	No multicollinearity
	Soil thickness	2.33	No multicollinearity
Environment	Distance to fault	1.22	No multicollinearity
	Distance to road	2.37	No multicollinearity
	NDVI	1.23	No multicollinearity
Conclusion	All values	<5	Independent

3.3. Non–Landslide Sampling Strategies

3.3.1. Strategy I: Improved Non–Landslide Sampling Strategy

Strategy I optimizes non–landslide sample selection by integrating physical slope stability analysis using TRIGRS model with 50 m buffering constraints. This approach aims to identify areas that are both physically stable and spatially distinct from existing landslides.

The TRIGRS model was employed to simulate the factor of safety (FoS) for slope grid cells under rainfall conditions [30]. The model calculates slope stability based on the infinite slope analysis method, expressed as:

$$F_s(Z,t) = \frac{\tan\varphi}{\tan\delta} + \frac{c - \psi(Z,t)\gamma_w \tan\varphi}{\gamma_s Z \sin\delta \cos\delta} \quad (2)$$

where c is effective cohesion, φ is the effective internal friction angle, γ_w is the unit weight of water, γ_s is the unit weight of soil, and $\Psi(Z, t)$ is the pressure head in the unsaturated layer calculated via Equation (3):

$$\psi(Z,t) = \frac{\cos\alpha}{\delta \cos^2\alpha} ln\left[\frac{K(Z,t)}{K_S}\right] + \psi_0 \quad (3)$$

where δ is the fitting parameter for the soil–water characteristic curve [31], α is the slope angle, $K(Z, t)$ is the unsaturated permeability coefficient, and Z is the soil thickness.

To implement the model, the parameterization for the five lithological types present in the study area is summarized in Table 3. Soil physical and mechanical parameters are derived from regional literature and geological reports [32]. Notably, as Siliceous rock (Crt) predominantly consists of exposed bedrock with minimal soil cover, rock mass parameters were applied to reflect its high stability, whereas soil parameters are used for other weathered lithologies. Regarding the hydraulic parameters, due to the lack of basin–wide in situ testing, representative empirical values were assigned based on the dominant soil properties in the region. Specifically, the saturated permeability (K_s) was set to 3.58×10^{-5} m/s, and the hydraulic diffusivity D_0 was initialized at 7.16×10^{-3} m^2/s. The saturated (θ_s) and residual (θ_r) water contents were set to 0.40 and 0.05, respectively [33]. To accurately initialize the hydraulic conditions, the model was initialized with a steady–state background infiltration rate (I_{Z0}) of 1.0×10^{-7} m/s, representing antecedent moisture conditions. The initial groundwater table was estimated spatially using a depth–dependent scaling factor ($d_w = 0.8Z$) based on regional hydrogeological surveys. Regarding the boundary conditions, the finite–depth analytical solution was employed (series expansion terms $N_{max} = 30$, $M_{max} = 20$). This configuration enforces a zero–flux (impermeable) boundary at the soil–bedrock interface to capture the formation of perched water tables, while the surface uses a flux boundary equal to the time–varying rainfall intensity.

https://doi.org/10.3390/rs18030408

Table 3. Soil physical and mechanical parameters of five lithological types used in TRIGRS model.

Parameter	Unit	Lithological Types				
		Mss	**Gr**	**Crt**	**Ls**	**Qss**
Cohesion (c')	KN/m^2	22	33.28	2340	800	247
Friction angle (φ')	°	37	41.37	35	40	40
Unit weight of soil (γ_s)	KN/m^3	26.8	18.2	27	26	24.9

While we acknowledge that assigning regional averages simplifies local heterogeneity caused by weathering or structural discontinuities, this approach provides a representative baseline for regional–scale screening. In terms of model configuration, the model was driven by a 72 h uniform rainfall intensity. A conservative stability threshold of FoS > 1.5 was adopted to define stable areas (referring to Table 4). This threshold aligns with standard geotechnical practices for permanent slopes [34], providing a safety margin to account for parameter uncertainties. Areas with FoS $\leq$ 1.5 were conservatively excluded to minimize false negatives.

Table 4. Classification of slope stability based on the factor of safety (FoS).

Stability	Unstable	Slightly Unstable	Basically Stable	Stable
Factor of safety	FoS $\leq$ 1.0	1.0 < FoS $\leq$ 1.25	1.25 < FoS $\leq$ 1.5	FoS > 1.5

The final non–landslide samples for Strategy I were selected from grid cells that satisfied two conditions: (1) being classified as stable by the TRIGRS model (FoS > 1.5), and (2) being located outside a 50 m buffer zone surrounding historical landslides. The 50 m constraint was applied to account for mapping uncertainties and to exclude potential tension cracks or unstable transition zones immediately adjacent to landslide bodies.

3.3.2. Strategy II: Traditional Buffering Method

Strategy II represents the traditional buffering method and serves as a benchmark. In this strategy, non–landslide samples were randomly selected based on a spatial constraint: they must be located outside a 50 m buffer zone of historical landslides [35]. This method relies on the assumption that areas spatially distant from known land–slides are stable, without considering the physical mechanical properties of the slope.

3.4. Machine Learning Models and Hyperparameter Optimization

To evaluate landslide susceptibility through diverse algorithmic perspectives, this study employs four representative machine learning models: multi–layer perceptron (MLP), random forest (RF), support vector machine (SVM), and extreme gradient boosting (XGBoost). These algorithms were specifically selected to cover the primary learning paradigms in current geospatial modeling: neural networks, bagging ensembles, kernel–based learning, and gradient boosting. By incorporating these distinct mechanisms, the study establishes a robust comparative baseline that transcends the limitations of any single algorithmic family [35–39].

To ensure optimal generalization and fair comparison, the hyperparameters for all models were rigorously tuned using the Bayesian optimization algorithm implemented via the optuna framework. The optimization process was configured as follows:

(1) Objective: minimize the validation loss to prevent overfitting.
(2) Search Strategy: the tree–structured parzen estimator (TPE) algorithm was used to efficiently explore the hyperparameter space (300 trials per model).

https://doi.org/10.3390/rs18030408

(3) Validation: a hold–out validation strategy (20% data split) was used within the optimization loop.

The optimal hyperparameters are summarized in Table 5.

Table 5. Hyperparameter search spaces and optimal settings for the four ML models.

ML Models	Hyperparameter	Search Space (Range)	Optimal Value
MLP	Hidden Layers	[3, 7]	3
	Learning Rate	$[5 \times 10^{-5}, 5 \times 10^{-2}]$ (log)	0.001
	Batch Size	[16, 128]	128
	Optimizer	[Adam, SGD]	Adam
SVM	Kernel	[RBF]	RBF
	C (Regularization)	[0.1, 100] (log)	1
	Gamma	[0.001, 1] (log)	Scale
RF	n_estimators	[100, 2000]	1000
	Max Depth	[5, 30]	15
	Max Features	[0.1, 0.9]	0.45
XGBoost	Learning Rate	[0.01, 0.3]	0.02
	Max Depth	[3, 15]	10
	Subsample	[0.5, 1.0]	0.95
	Colsample_bytree	[0.5, 1.0]	0.73

3.4.1. Multi–Layer Perceptron Model

MLP is a feedforward neural network capable of capturing complex non–linear relationships [40]. In this study, the MLP classifier architecture was rigorously optimized via Bayesian optimization (Optuna), resulting in a structure comprising three primary fully connected hidden layers. The first hidden layer consists of 832 neurons with GELU activation and a dropout rate of 0.50, followed by batch normalization. The second hidden layer contains 416 neurons with GELU activation and a dropout rate of 0.30. The third hidden layer comprises 416 neurons with Leaky ReLU activation, a dropout rate of 0.10, and batch normalization. Additionally, a dense layer with 32 neurons (ReLU activation, dropout 0.20) was implemented before the final sigmoid output layer. The model was trained using the RMSprop optimizer with a learning rate of 0.0015 and a batch size of 128.

3.4.2. Support Vector Machine Model

SVM classifies data by constructing an optimal hyperplane in a high–dimensional space [41]. Based on the optimization results, the radial basis function (RBF) kernel was selected with a regularization parameter C = 1, balancing the trade–off between margin maximization and classification error.

3.4.3. Random Forest Model

RF is a bagging–based ensemble algorithm that aggregates predictions from multiple decision trees to reduce variance [42]. The optimal configuration for this dataset included an ensemble of 1000 trees with a maximum depth of 15 and a feature sampling ratio of 0.45.

3.4.4. Extreme Gradient Boosting Model

XGBoost is an efficient implementation of gradient boosting designed for high performance with high–dimensional data [43]. The model was tuned with a learning rate of 0.02 and a maximum tree depth of 10. To prevent overfitting, subsample and column sampling ratios were set to 0.95 and 0.73, respectively.

https://doi.org/10.3390/rs18030408

3.5. Validation Method

3.5.1. Receiver Operating Characteristics

The model performance was primarily evaluated using the area under the receiver operating characteristic curve (AUC) [44]. The ROC curve plots the true positive rate (sensitivity) against the false positive rate (1–specificity) at various thresholds. AUC values range from 0.5 to 1, where values exceeding 0.9 are generally considered to indicate excellent predictive performance.

3.5.2. Confusion Matrix

To provide a comprehensive assessment, statistical metrics derived from the confusion matrix were calculated, including accuracy, precision, recall, and F1–score [45]. Accuracy (ACC) represents the proportion of correctly classified samples and is calculated as:

$$ACC = \frac{TP + TN}{TP + FN + FP + TN} \tag{4}$$

where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively.

3.5.3. Random and Spatial Cross–Validation

This study employs both random and spatial cross–validation to rigorously evaluate model performance. Random 5–fold cross–validation was used for hyperparameter optimization and baseline assessment, while spatial cross–validation was adopted to evaluate the model's transferability to unseen areas. Model performance was quantified using ROC curves and confusion matrices.

Random 5–fold cross–validation involves randomly partitioning the entire dataset into five equal subsets, iteratively using four for training and one for validation. While effective for baseline testing, this method ignores spatial dependencies. To address this, we implemented spatial cross–validation based on administrative boundaries. The study area was divided into five distinct towns. In each iteration, data from four towns served as the training set, while the remaining town functioned as a spatially independent testing set. By strictly separating training and testing samples geographically, this method robustly assesses the model's ability to predict landslides in unseen regions.

4. Results

4.1. Landslide Inventories

The landslides inventory database in the study area was established through manual visual interpretation, analyzing alterations in image color tones and topographic features in pre– and post–event high–resolution satellite imageries and digital elevation models (DEMs) (Figure 5a,b). To ensure the reliability of the database and filter out noise, the minimum mapping unit (MMU) was set to 50 m^2 (approximately 200 pixels in the post–event imagery). For the sampling strategy, the highest point of the scarp (crown area) was extracted for each landslide polygon. This specific location was selected to characterize the initiation zone as its topographical and geological conditions at the scarp are little difference. Hence, positive samples could be taken from the highest point of landslides [46].

https://doi.org/10.3390/rs18030408

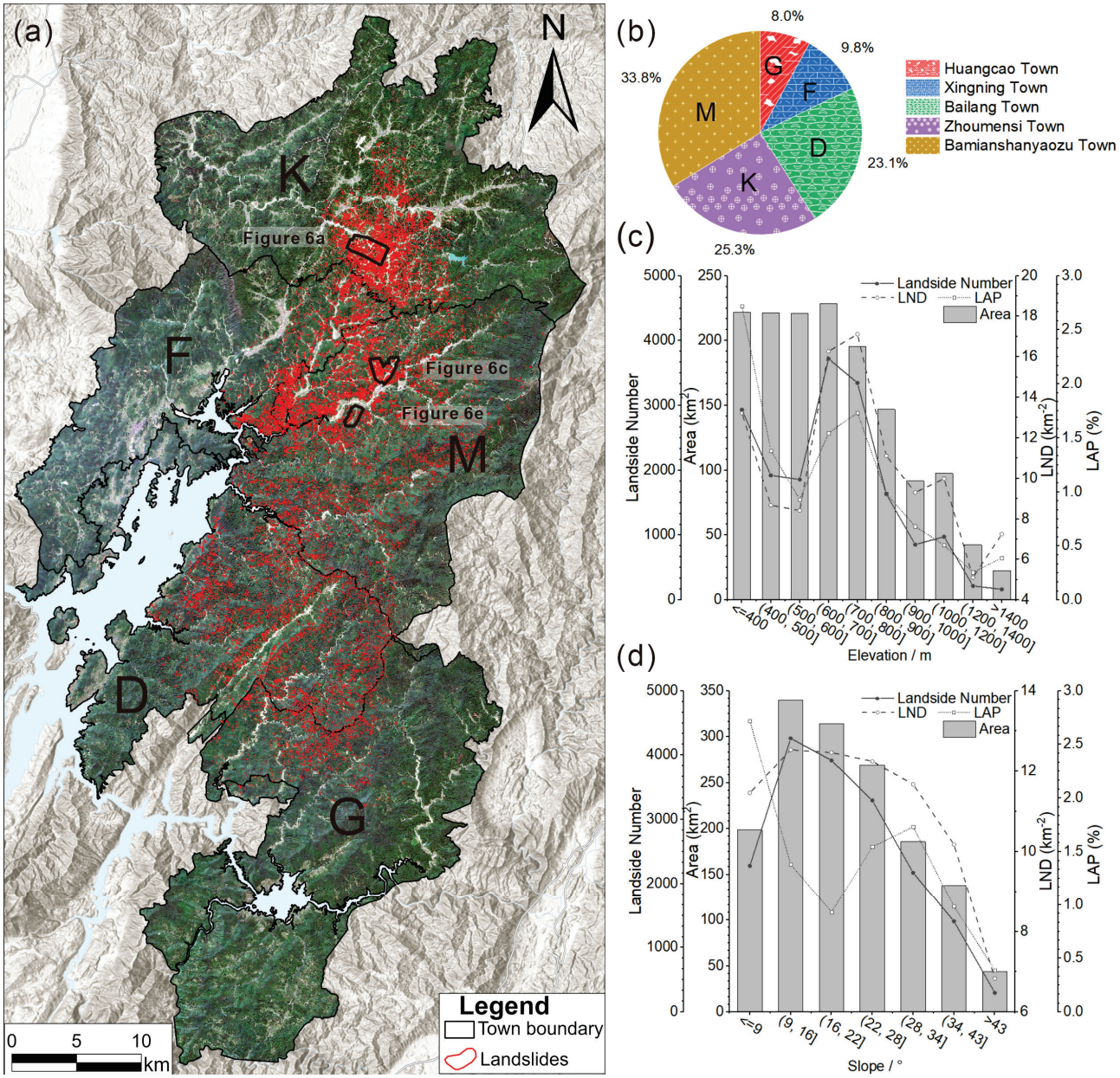


Figure 5. Spatial distribution and statistics of the landslide inventory. (**a**) Spatial clustering of landslides triggered by the extreme rainstorm event; (**b**) Landslide reporting statistics for towns within the study area; (**c**) Statistical analysis of landslides and elevation; (**d**) Statistical analysis of landslides and slope.

The 2024 Typhoon Gaemi triggered a total of 17,529 landslides in Zixing County (Figure 5a). From a geographical perspective, landslides are predominantly concentrated in Bamianshanyaozu Town and Zhoumensi Town. Field investigations and UAV images (resolution with 0.05 m) were conducted to validate the database's accuracy. Based on the verification results for three representative areas (Figure 6) and the high resolution of remote sensing imagery (0.7 m), the positional accuracy is estimated to be within 1 m, and the interpretation accuracy exceeds 90%.

https://doi.org/10.3390/rs18030408

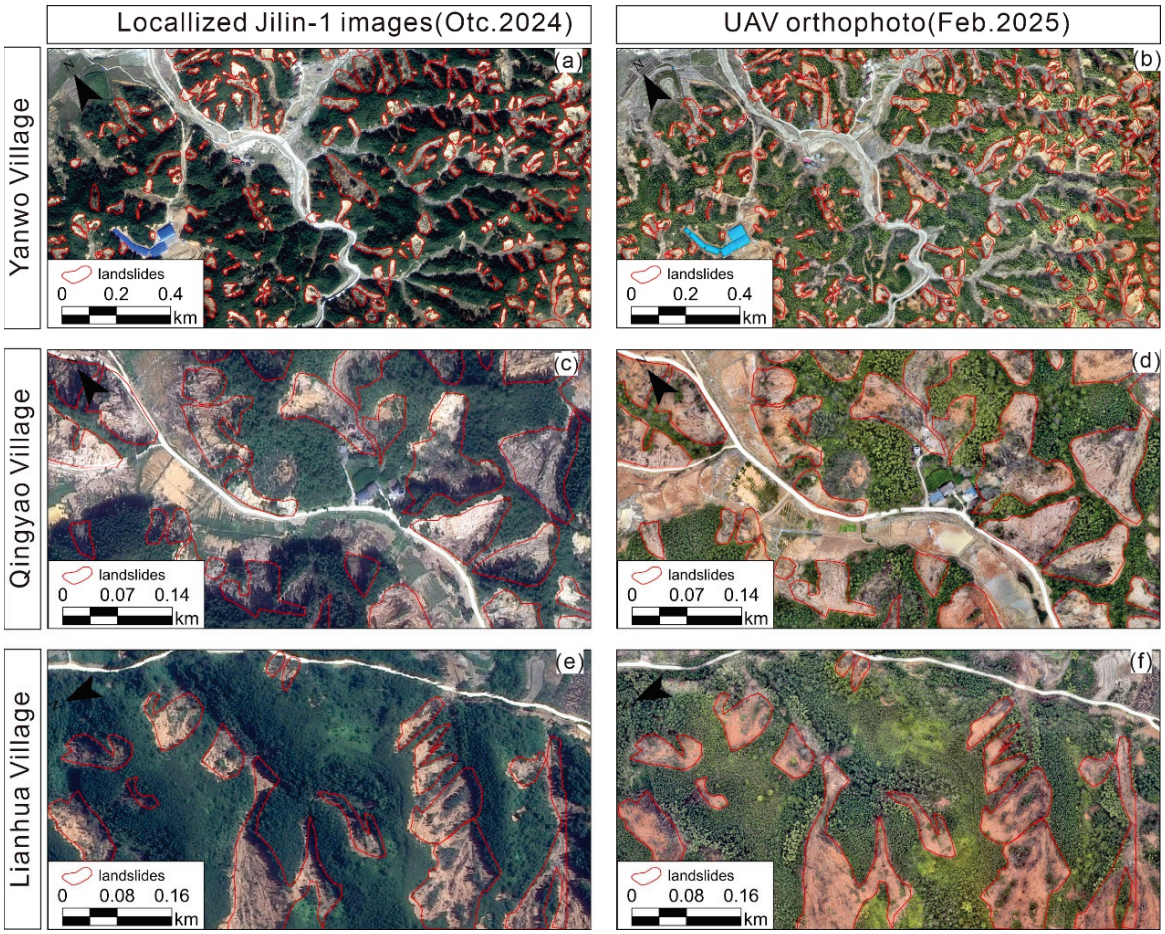


Figure 6. Validation of the landslide inventory using UAV images in three typical villages. (**a**,**b**) Pre– and post–event interpretation in Yanwo Village; (**c**,**d**) Pre– and post–event interpretation in Qingyao Village; (**e**,**f**) Pre– and post–event interpretation in Lianhua Village.

This study adopts two widely used indicators—landslide number density (LND), defined as the number of landslides per square kilometer, and landslide area percentage (LAP), defined as the ratio of landslide area to the total area of each factor category—to analyze the spatial patterns of rainfall–induced landslides and their relationships with controlling factors (Figure 5c,d) [47]. By integrating LND, LAP, and the absolute landslide count, we conduct a comprehensive assessment of landslide development intensity across different factor classes. The interpretation of these metrics reveals distinct scenarios: high LND with low LAP indicates numerous but small–scale landslides; low LND with high LAP suggests fewer but large–area failures; and concurrent high values of both LND and LAP signify severe overall landslide activity. Notably, the magnitude of LND and LAP is influenced by the size of classification intervals, underscoring the necessity of combining these indices with landslide counts for a robust and balanced evaluation.

Elevation of Zixing County ranges from 268 m to 2042 m. The elevation was classified into intervals of 100 m (for elevations below 1000 m) and 200 m (for higher elevations) to analyze the landslide distribution. Statistical results indicate that a total of 15,417 landslides (87.9% of all recorded events) occurred at elevations $\leq$ 900 m. As shown in Figure 5c, both landslide number density (LND) and landslide area percentage (LAP) exhibit a non–monotonic pattern with elevation—initially decreasing, then increasing, and finally decreasing again—with the peak LND and LAP values occurring in the (700, 800] m interval. Regarding slope, the maximum gradient in Zixing reaches 89°.

https://doi.org/10.3390/rs18030408

The slope gradient was classified into seven intervals using the natural breaks (Jenks) method to reflect the natural grouping of data values. Statistical analysis reveals that 15,924 landslides (90.8% of the total) occurred on slopes $\leq 34°$. Analysis of Figure 5d reveals that LND increases with slope up to a maximum in the (9°, 16°] interval and then declines, whereas LAP first decreases, then increases, and subsequently decreases again, peaking at slopes $\leq 9°$.

4.2. Distribution of Non–Landslide Samples

Analysis of the factor of safety (FoS) distribution map produced by the TRIGRS model reveals that 14,270 of the recorded landslides (81%) are concentrated in unstable areas with an FoS below 1.5, whereas only 3347 landslide points (19%) occur in highly stable zones (FS > 1.5) (Figure 7a,c). This pronounced spatial alignment between landslide occurrences and low–FoS regions underscores the physical credibility of using TRIGRS outputs to guide non–landslide sample selection. To minimize bias arising from mapping uncertainties and spatial autocorrelation, a supplementary 50 m buffer zone is excluded around each landslide perimeter, yielding the final set of non–landslide samples used for model development (Figure 7b,d).

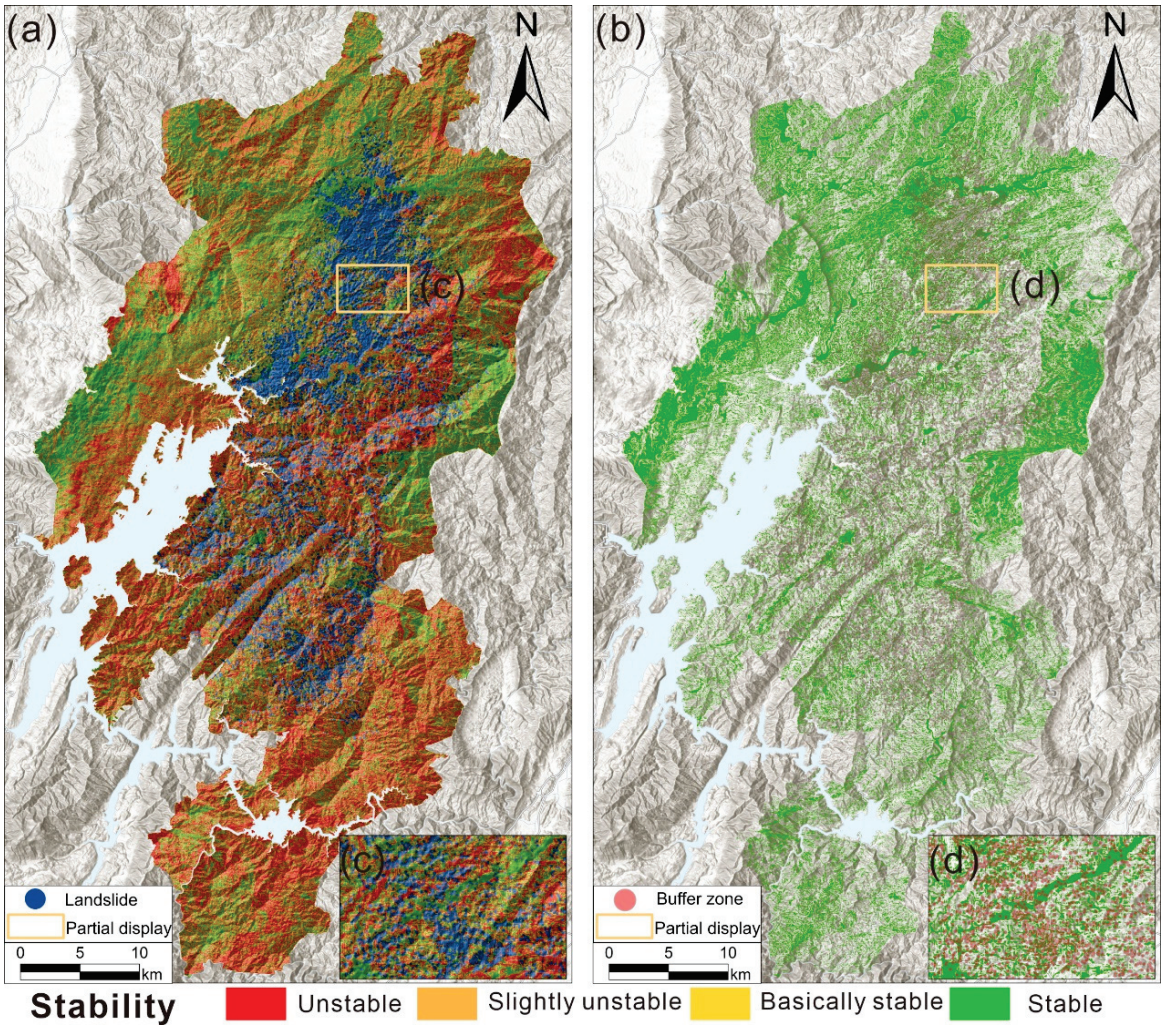


Figure 7. Illustration of the improved non–landslide sampling strategy (Strategy I). (**a**) Factor of safety (FoS) in the study area; (**b**) Areas with FoS > 1.5 and located outside landslide buffer zones; (**c**) Detailed view showing the overlay of FoS values on landslide locations; (**d**) Detailed view of areas with FoS > 1.5 overlaid on landslide buffer zones.

https://doi.org/10.3390/rs18030408

4.3. Landslide Susceptibility Mapping

Landslide susceptibility maps are produced using four machine learning algorithms: multilayer perceptron (MLP), random forest (RF), support vector machine (SVM), and XGBoost. The number of grid cells corresponding to each susceptibility level for different models under strategy I is presented in Table 6.

Table 6. The number of grid cells at each susceptibility level for the four models under sampling strategy I.

Sampling Strategy	LSM Class				
	Very Low	Low	Medium	High	Very High
MLP–Strategy I	269,343	28,468	36,111	40,826	60,169
RF–Strategy I	150,643	115,865	94,983	51,003	22,423
SVM–Strategy I	156,116	108,768	94,472	50,106	25,455
XGBoost–Strategy I	222,863	65,797	64,644	49,491	32,122

The susceptibility maps are classified into five levels based on index values: 0–0.2 (very low), 0.2–0.4 (low), 0.4–0.6 (medium), 0.6–0.8 (high), and 0.8–1.0 (very high) (Figure 8). Given that an effective model should assign high probabilities to actual landslide locations, we focused our evaluation on the cumulative proportion of landslides falling within the high to very high susceptibility classes (0.6–1.0).

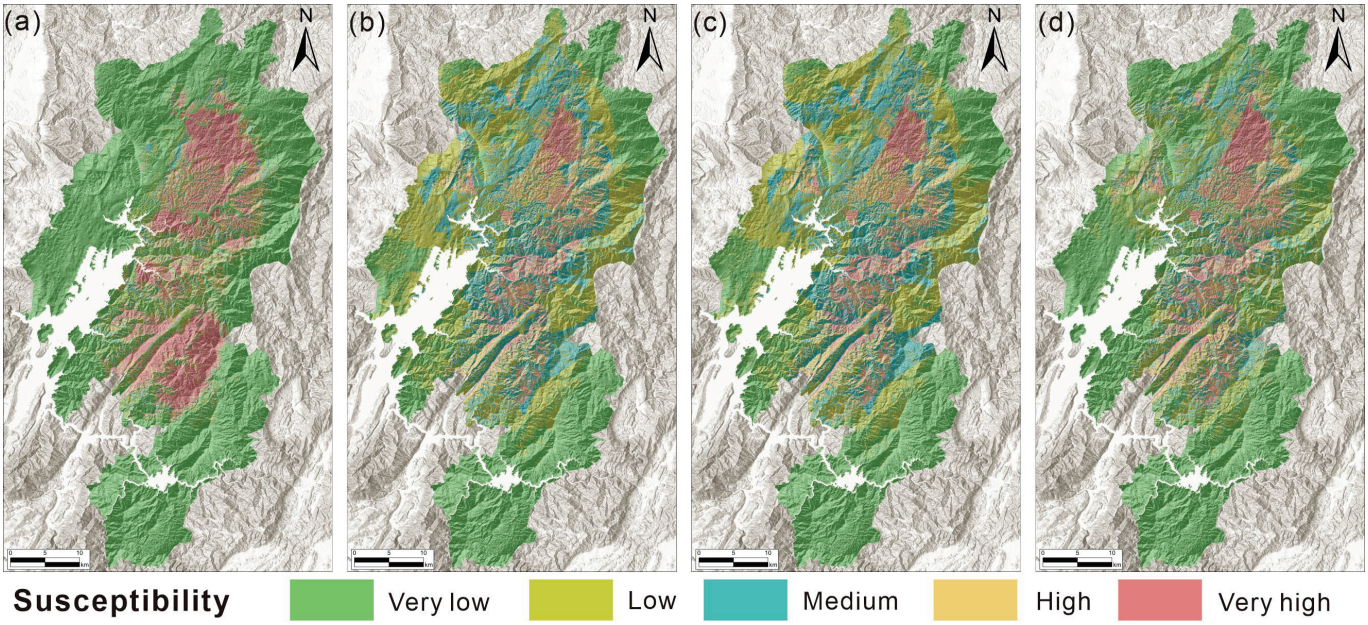


Figure 8. Landslide susceptibility maps generated by four machine learning models using the improved non–landslide sampling strategy (Strategy I). (**a**) MLP model; (**b**) RF model; (**c**) SVM model; (**d**) XGBoost model.

The analysis reveals that the capture rates for the RF, SVM, and XGBoost models in this critical range were relatively limited, accounting for only 62.1%, 62.7%, and 65.4% of the historical landslides, respectively. In distinct contrast, the MLP model demonstrated superior sensitivity, successfully predicting 90.0% of the inventory landslides within these high–risk regions. This represents a substantial performance margin of 24.6% to 27.9% over the benchmark models. These findings indicate that the MLP model generates a probability distribution that aligns more closely with the actual spatial distribution of landslides, effectively concentrating high–risk predictions on unstable slopes (Figure 9).

https://doi.org/10.3390/rs18030408

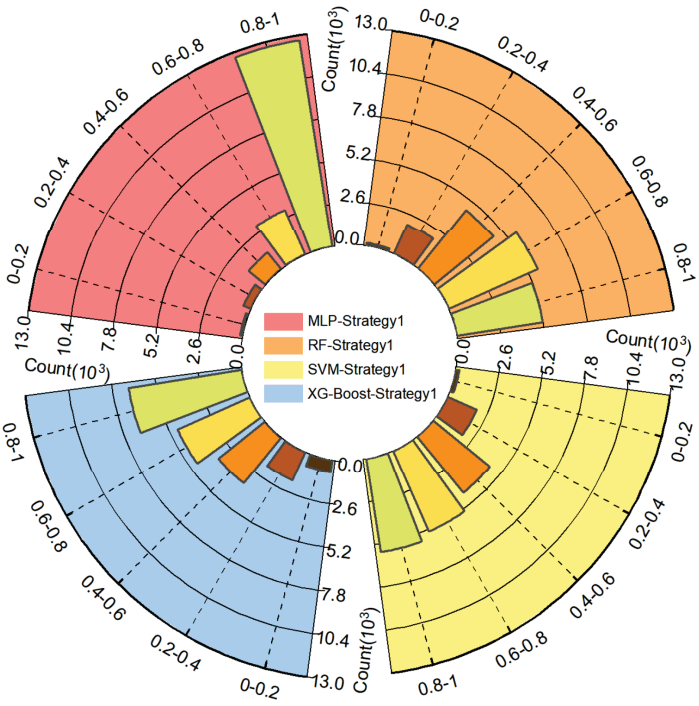


Figure 9. Distribution of landslides across different susceptibility zones under Strategy I.

4.4. Comparison of Model Performance

To evaluate both the learning capability and the spatial transferability of the models, we conducted a comprehensive performance assessment using two distinct validation protocols: (1) random 5–fold cross–validation, which establishes the baseline model accuracy; and (2) spatial cross–validation, which tests the model's ability to generalize to unseen geographic regions. The results from these two approaches are presented below.

4.4.1. Random Cross–Validation

Under random 5–fold cross–validation, the performance of the four machine learning models was evaluated using ROC curves (Figure 10a) and detailed statistical metrics (Table 7). All models performed robustly with sampling strategy I, with MLP model achieving the highest mean AUC of 0.934, significantly outperforming XGBoost (0.907), RF (0.902), and SVM (0.886). Paired *t*–tests confirmed these differences were statistically significant ($p < 0.001$). MLP model also achieved the highest accuracy (0.86) and F1–score (0.87), demonstrating strong capability in handling the complex non–linear relationships inherent in landslide susceptibility modeling (Table 8).

Table 7. Quantitative comparison of model performance (AUC) under sampling strategy I using random 5–fold cross–validation and paired *t*–tests.

Fold	MLP (Proposed)	RF	SVM	XGBoost
Fold 1	0.932	0.900	0.879	0.905
Fold 2	0.933	0.901	0.883	0.910
Fold 3	0.935	0.901	0.887	0.899
Fold 4	0.937	0.903	0.890	0.915
Fold 5	0.931	0.905	0.891	0.906
Mean AUC	0.934	0.902	0.886	0.907
Std. dev	±0.002	±0.002	±0.005	±0.006
p–value (vs. MLP)	–	<0.001	<0.001	<0.001

https://doi.org/10.3390/rs18030408

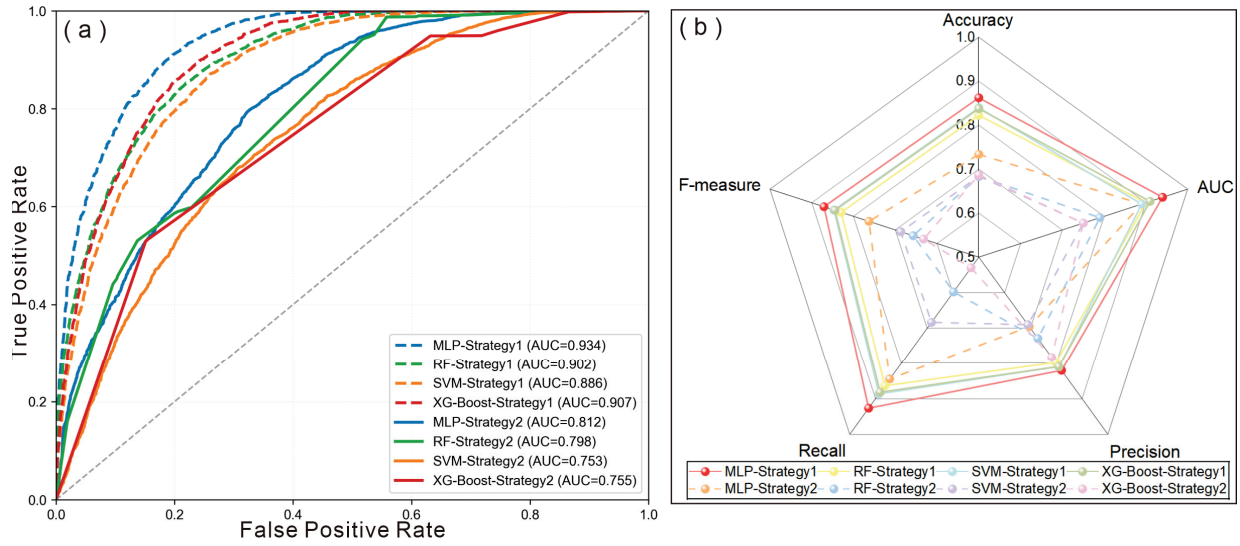


Figure 10. Performance evaluation of eight susceptibility models. (**a**) ROC curves of eight models; (**b**) Radar chart showing secondary performance metrics for the eight models.

Table 8. Detailed performances of four machine learning models under two sampling strategies.

Model	Strategy	AUC	Accuracy	Precision	Recall	Specificity	F1–Score
MLP	Strategy I	0.934	0.862	0.821	0.927	0.798	0.871
	Strategy II	0.812	0.734	0.695	0.845	0.619	0.763
RF	Strategy I	0.902	0.823	0.798	0.864	0.781	0.830
	Strategy II	0.798	0.683	0.729	0.598	0.771	0.654
SVM	Strategy I	0.886	0.840	0.810	0.887	0.792	0.847
	Strategy II	0.753	0.684	0.690	0.683	0.684	0.687
XGBoost	Strategy I	0.907	0.837	0.810	0.882	0.793	0.844
	Strategy II	0.755	0.687	0.783	0.530	0.849	0.632

To validate the effectiveness of the proposed Strategy I, we compared it against Strategy II (traditional buffering method) across four machine learning algorithms. To quantify the performance enhancement, the relative improvement rate for each metric (e.g., AUC, accuracy, specificity) was calculated as:

$$Improvement = \frac{Metric_{StrategyI} - Metric_{StrategyII}}{Metric_{StrategyII}} \times 100\% \quad (5)$$

Under random cross–validation, Strategy I consistently improved AUC across all models. For the MLP model, AUC values significantly increased from 0.812 to 0.934, representing a 15.02% improvement (Figure 10). Similarly, the RF, SVM, and XGB models exhibited AUC gains of 13.03%, 17.66%, and 20.13%. Meanwhile, the average improvement in AUC across the four models is 16.46%.

For the optimal MLP model, specificity improved from 61.9% (strategy II) to 79.8% (strategy I) (Table 8), indicating a significant reduction in false positives. This improvement stems from Strategy I's ability to filter out physically unstable areas mislabeled as non–landslides in Strategy II [48,49]. Consequently, strategy I correctly classified 2812 non–landslide samples, an increase of 667 over the 2145 identified under strategy II. The improved non–landslide sampling strategy (strategy I) enhanced the predictive accuracy in comparison with the traditional buffering method (Strategy II) (Figure 11). This performance improvement is primarily attributed to the integration of physical constraints, which enhances data reliability by filtering out physically unstable areas that traditional

https://doi.org/10.3390/rs18030408

buffering method overlooks. By employing the TRIGRS model to exclude these ambiguous samples, which may be geologically unstable despite lacking historical failures [50,51], Strategy I prevents the inclusion of noise that would otherwise disrupt classification boundaries, thereby ensuring high–quality training samples [52,53].

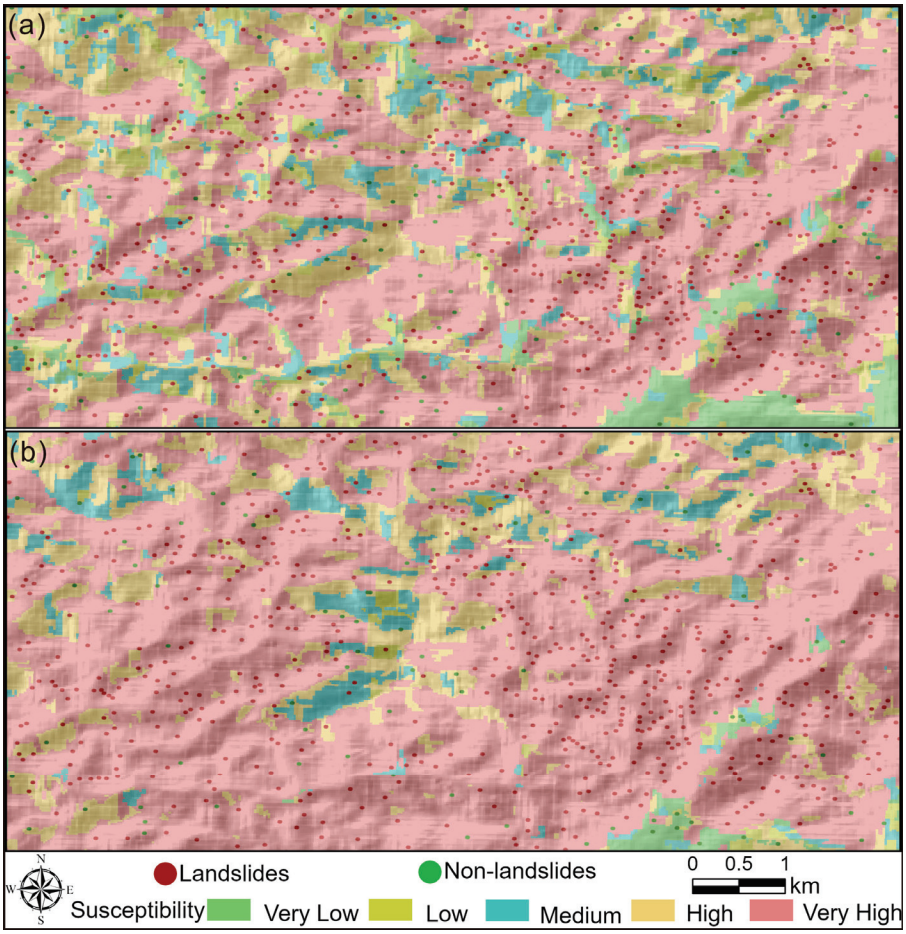


Figure 11. Comparison of two non–landslide sampling methods for constructing MLP–based susceptibility maps. (**a**) Strategy I: Non–landslide samples are selected from physically stable areas (FoS > 1.5) excluding landslide buffer zones; (**b**) Strategy II: traditional buffering method where non–landslide samples are selected randomly from any area outside the landslide buffer zones, without physical stability constraints. (Locations as shown in Figure 7c).

4.4.2. Spatial Cross–Validation

Results under spatial cross–validation further confirmed the generalizability of Strategy I (Table 9). Across five spatially independent administrative regions, Strategy I achieved a mean AUC of 0.854, outperforming Strategy II by 29.66%. It is worth noting that the mean AUC decreased from 0.934 (random CV) to 0.854 (spatial CV). This performance drop confirms that the administrative–based partitioning effectively reduced spatial autocorrelation and information leakage, suggesting that the test towns possessed sufficient geological and topographical heterogeneity to rigorously challenge the model's generalizability. This pronounced improvement underscores the particular efficacy of Strategy I in addressing the spatial overfitting that commonly affects traditional buffering method. Critically, Strategy I effectively mitigated the high false–positive rate observed in Strategy II. For instance, in Town M, specificity improved dramatically from 0.163 under Strategy II to 0.891 under Strategy I. These findings confirm that the physics–integrated strategy not only enhances global accuracy but also significantly improves transferability to spatially independent areas.

https://doi.org/10.3390/rs18030408

Table 9. Spatial cross–validation results of the MLP model under two sampling strategies.

Test Region (Town)	Strategy I AUC	Strategy I Accuracy	Strategy I Specificity	Strategy II AUC	Strategy II Accuracy	Strategy II Specificity	Improvement (AUC)
K	0.799	0.622	0.830	0.589	0.539	0.589	35.65%
F	0.767	0.631	0.774	0.636	0.459	0.733	20.60%
G	0.908	0.855	0.931	0.585	0.494	0.322	55.21%
D	0.879	0.754	0.634	0.762	0.608	0.348	15.35%
M	0.916	0.788	0.891	0.754	0.547	0.163	21.49%
Mean ± Std.	0.854 ± 0.067	0.730 ± 0.101	0.812 ± 0.116	0.665 ± 0.087	0.529 ± 0.057	0.431 ± 0.227	29.66% ± 16.14%

5. Discussion

5.1. The Mechanism of Strategy I in Enhancing Model Performance

The SHAP analysis was employed to interpret the global importance of influencing factors. The importance ranking was determined based on the mean absolute SHAP value, which represents the average marginal contribution of each feature to the model's prediction output. As shown in Figure 12, rainfall ranks first with a mean absolute SHAP value of 0.145, followed closely by groundwater table (0.144) and distance to road (0.084). Quantitative analysis reveals that the cumulative contribution of the top six factors (rainfall, groundwater table, distance to road, aspect, distance to river, and distance to fault) accounts for approximately 80.1% of the total feature importance. This finding indicates that heavy precipitation and hydrological conditions are the primary triggers of landslides in the study area (Figure 13a,b), while anthropogenic activities (distance to roads) and geological structures also exert significant influence [54].

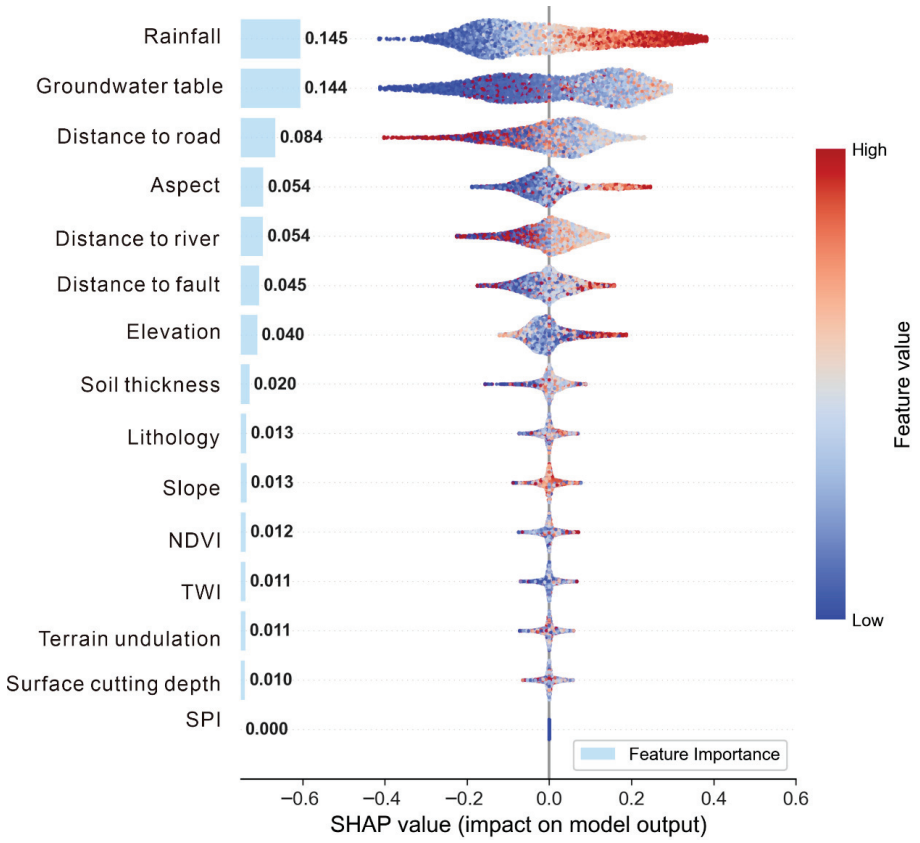


Figure 12. Feature importance from SHAP analysis.

https://doi.org/10.3390/rs18030408

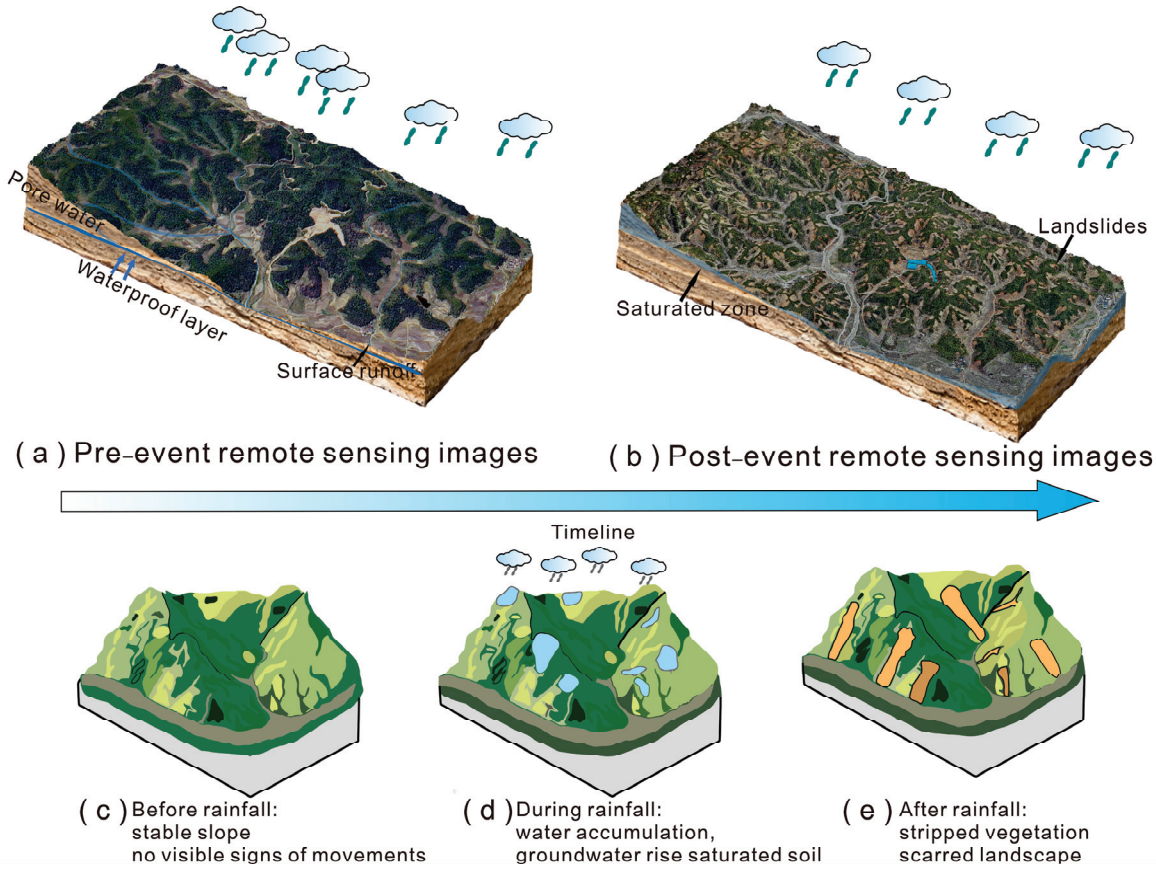


Figure 13. Schematic diagram of the physical mechanism driving rainfall–induced landslides. (**a**) pre–event remote sensing images; (**b**) post–event remote sensing images; (**c**) before rainfall stable slope no visible signs of movements; (**d**) during rainfall water accumulation, groundwater rise saturated soil; (**e**) after rainfall stripped vegetation scarred landscape.

The clustered landslides triggered by Typhoon Gaemi in 2024 evolved through three distinct failure stages [55,56] (Figure 13c,e): (a) initial stage with favorable drainage conditions and thinner accumulation mass; (b) as rainfall infiltrates the slope, rising soil moisture content gradually undermines slope stability; (c) slope failure stage with continuous rainfall saturating the unsaturated zone, significantly reducing the slope shear strength. Critical physical and mechanical processes in the TRIGRS model include rainfall infiltration and groundwater dynamics, consistent with key factors from SHAP analysis. This alignment suggests that the strategy I effectively captures the fundamental physical processes of slope instability. Mutual validation between physical mechanisms and data–driven approach provides compelling evidence for the landslide failure mechanisms. Furthermore, the landslide susceptibility assessment coupling physics–based model with machine learning significantly enhances model interpretability and prediction reliability.

5.2. Comparative Analysis of SHAP Importance and Statistical Correlation

To further investigate the model's interpretability, we compared the SHAP feature importance with the Pearson correlation analysis (Figure 12). Discrepancies between the two metrics highlight the MLP model's advantage in capturing complex non–linear relationships that traditional linear statistics may overlook.

Firstly, regarding factors with low correlation but high importance, aspect and distance to road are prime examples. Although aspect typically shows a weak linear correlation with landslide distribution due to its cyclical nature (0–360°), it ranks 4th in SHAP importance. This indicates that the model successfully captured the windward slope effect of Typhoon

https://doi.org/10.3390/rs18030408

Gaemi, where specific slope orientations received significantly more precipitation and wind load. Similarly, distance to road (rank 3) exhibits a threshold–based non–linear impact and slope stability is compromised mainly in the immediate vicinity of road cuts, a pattern is often underestimated by linear correlation coefficients.

Secondly, regarding factors that appear highly correlated or physically important but ranked lower, such as TWI (rank 12) and SPI (rank 15), this can be attributed to the model's handling of information redundancy. As shown in the correlation matrix (Figure 6), these morphological indices share information with slope and elevation. Critically, since the model already identified rainfall (rank 1) and groundwater table (rank 2) as the dominant features, the marginal contribution of static hydrological proxies like TWI was reduced. This suggests that the physics–informed sampling strategy (strategy I) guides the model to prioritize direct physical drivers (dynamic hydrology) over indirect static indicators.

6. Conclusions

This study proposes an improved non–landslide sampling strategy for landslide susceptibility assessment. The performance of the improved non–landslide sampling strategy was evaluated by comparing the ROC values across four machine learning models. The key findings of this study are summarized below:

(1) The improved non–landslide sampling strategy, which integrates TRIGRS–derived stability (FoS > 1.5) with 50 m buffering constraints, significantly improved model performance over traditional buffering method. Using a 0.5 classification threshold, the improved strategy yielded substantial performance gains, with average improvements of 16.46% in AUC, 9.75% in specificity, and 16.78% in overall accuracy across all models.

(2) Under the improved non–landslide sampling strategy, the MLP model outperformed RF, SVM, and XGBoost, achieving an AUC of 0.934 and an accuracy of 86.2%.

(3) SHAP analysis identified rainfall, groundwater table, and human engineering activities as dominant factors controlling landslide initiation during Typhoon Gaemi. This confirms that the improved strategy effectively captures the fundamental mechanisms of slope instability.

Future work could aim to: (1) integrate high–precision deformation monitoring data for identifying actively unstable slopes; and (2) refine input parameters for physical simulations using probabilistic methods to quantify uncertainties and improve model generalization.

Author Contributions: Conceptualization, M.C. and P.L.; methodology, D.P., M.C., P.L., S.X. and Y.S.; software, M.C. and Y.S.; validation, M.C., Y.Z. and Q.X.; formal analysis, D.P.; investigation, M.C.; data curation, D.P., Y.Z. and Y.S.; writing—original draft preparation, D.P. and M.C.; writing—review and editing, D.P., M.C., Y.Z., P.L., S.X., B.T. and L.K.; project administration, D.P. and Q.X.; funding acquisition, D.P. and Q.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key R&D Program of China (No. 2022YFC3003205), the National Natural Science Foundation of China (No. 42377196), the Sichuan Science and Technology Program (No. 2024NSFSC1997), Chengdu University of Technology Postgraduate Innovative Cultivation Program (No. 2024BJCX009), State Key Laboratory of Geohazard Prevention and Geoenvironment Protection Independent Research Project (No. SKLGP2022Z028), and Joint Research Foundation of Gansu Province (Nos. 24JRRA800 and 25JRRA1158).

Data Availability Statement: Data will be made available on request.

Acknowledgments: Thanks to Caiyun Technology Company for providing us with the rainfall radar data. We also thank the editor and all reviewers for their valuable comments.

https://doi.org/10.3390/rs18030408

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

Gr	Granite
Mss	Metasandstone
Ctr	Siliceous rock
Ls	Limestone
Qss	Quartz sandstone
MLP	Multi–layer Perceptron
RF	Random Forest
SVM	Support Vector Machine
XGBoost	Extreme Gradient Boosting
AUC	Area Under the Curve
SHAP	SHapley Additive exPlanations
NDVI	Normalized Difference Vegetation Index
FoS	Factor of safety
RBF	Radial Basis Function
ROC	Receiver Operating Characteristic
TPR	True positive rate
FPR	False positive rate
TPs	True Positives
TNs	True Negatives
FPs	False Positives
FNs	False Negatives
ACC	Accuracy
TRIGRS	Transient Rainfall Infiltration and Grid–based Regional Slope–Stability Model

References

1. Iverson, R. Landslide triggering by rain infiltration. *Water Resour. Res.* **2000**, *36*, 1897–1910. [CrossRef]
2. Shan, K.; Lin, Y.; Chu, P.-S.; Yu, X.; Song, F. Seasonal advance of intense tropical cyclones in a warming climate. *Nature* **2023**, *623*, 83–89. [CrossRef]
3. Tollefson, J. Severe weather linked more strongly to global warming. *Nature* **2015**, *520*, 20. [CrossRef]
4. Zhao, H.; Xu, Q.; Chen, W.; Xu, F.; Peng, D.; Pu, C.; Liu, R.; Shen, Y. Improving rainfall-triggered landslide susceptibility mapping through source-area boundary sampling and multi-dimensional feature analysis. *Gondwana Res.* **2025**, *153*, 265–283. [CrossRef]
5. Kong, L.; Feng, W.; Yi, X.; Xue, Z.; Bai, L. Enhanced landslide susceptibility mapping in data-scarce regions via unsupervised few-shot learning. *Gondwana Res.* **2025**, *138*, 31–46. [CrossRef]
6. Mirus, B.B.; Jones, E.S.; Baum, R.L.; Godt, J.W.; Slaughter, S.; Crawford, M.M.; Lancaster, J.; Stanley, T.; Kirschbaum, D.B.; Burns, W.J.; et al. Landslides across the USA: Occurrence, susceptibility, and data limitations. *Landslides* **2020**, *17*, 2271–2285. [CrossRef]
7. Tsangaratos, P.; Ilia, I.; Hong, H.; Chen, W.; Xu, C. Applying Information Theory and GIS-based quantitative methods to produce landslide susceptibility maps in Nancheng County, China. *Landslides* **2017**, *14*, 1091–1111. [CrossRef]
8. Huang, W.; Ding, M.; Li, Z.; Yu, J.; Ge, D.; Liu, Q.; Yang, J. Landslide susceptibility mapping and dynamic response along the Sichuan-Tibet transportation corridor using deep learning algorithms. *Catena* **2023**, *222*, 106866. [CrossRef]
9. Yang, C.; Liu, L.-L.; Huang, F.; Huang, L.; Wang, X.-M. Machine learning-based landslide susceptibility assessment with optimized ratio of landslide to non-landslide samples. *Gondwana Res.* **2023**, *123*, 198–216. [CrossRef]
10. Reichenbach, P.; Rossi, M.; Malamud, B.D.; Mihir, M.; Guzzetti, F. A review of statistically-based landslide susceptibility models. *Earth-Sci. Rev.* **2018**, *180*, 60–91. [CrossRef]
11. Merghadi, A.; Yunus, A.P.; Dou, J.; Whiteley, J.; ThaiPham, B.; Bui, D.T.; Avtar, R.; Abderrahmane, B. Machine learning methods for landslide susceptibility studies: A comparative overview of algorithm performance. *Earth-Sci. Rev.* **2020**, *207*, 103225. [CrossRef]
12. Kudaibergenov, M.; Nurakynov, S.; Iskakov, B.; Iskaliyeva, G.; Maksum, Y.; Orynbassarova, E.; Akhmetov, B.; Sydyk, N. Application of artificial intelligence in landslide susceptibility assessment: Review of recent progress. *Remote Sens.* **2024**, *17*, 34. [CrossRef]

https://doi.org/10.3390/rs18030408

13. Zhang, L.; Guo, Z.; Qi, S.; Zhao, T.; Wu, B.; Li, P. Landslide susceptibility evaluation and determination of critical influencing factors in eastern Sichuan mountainous area, China. *Ecol. Indic.* **2024**, *169*, 112911. [CrossRef]
14. Youssef, A.M.; Pourghasemi, H.R. Landslide susceptibility mapping using machine learning algorithms and comparison of their performance at Abha Basin, Asir Region, Saudi Arabia. *Geosci. Front.* **2021**, *12*, 639–655. [CrossRef]
15. Chang, Z.; Catani, F.; Huang, F.; Liu, G.; Meena, S.R.; Huang, J.; Zhou, C. Landslide susceptibility prediction using slope unit-based machine learning models considering the heterogeneity of conditioning factors. *J. Rock Mech. Geotech. Eng.* **2023**, *15*, 1127–1143. [CrossRef]
16. Wei, X.; Zhang, L.; Gardoni, P.; Chen, Y.; Tan, L.; Liu, D.; Du, C.; Li, H. Comparison of hybrid data-driven and physical models for landslide susceptibility mapping at regional scales. *Acta Geotech.* **2023**, *18*, 4453–4476. [CrossRef]
17. Wang, J.; Wang, Y.; Li, M.; Qi, Z.; Li, C.; Qi, H.; Zhang, X. Improved landslide susceptibility assessment: A new negative sample collection strategy and a comparative analysis of zoning methods. *Ecol. Indic.* **2024**, *169*, 112948. [CrossRef]
18. Liu, L.; Duan, C.; Gao, J.; Xiao, H.; Zhu, W.; Yang, C. Landslide susceptibility assessment using machine learning with a novel SHAP-based sampling strategy. *Geosci. Front.* **2025**, 102188. [CrossRef]
19. Lyu, H.-M.; Yin, Z.-Y.; Hicher, P.-Y.; Laouafa, F. Incorporating mitigation strategies in machine learning for landslide susceptibility prediction. *Geosci. Front.* **2024**, *15*, 101869. [CrossRef]
20. Deng, Z.; Lan, H.; Li, L.; Liu, Y.; Tian, N. A catchment-scale landslide hydro-mechanical coupling model considering spatial heterogeneity. *J. Hydrol.* **2025**, 134855. [CrossRef]
21. Hong, H.; Wang, D.; Zhu, A.-X.; Wang, Y. Landslide susceptibility mapping based on the reliability of landslide and non-landslide sample. *Expert Syst. Appl.* **2024**, *243*, 122933. [CrossRef]
22. Zhou, X.; Wen, H.; Zhang, Y.; Xu, J.; Zhang, W. Landslide susceptibility mapping using hybrid random forest with GeoDetector and RFE for factor optimization. *Geosci. Front.* **2021**, *12*, 101211. [CrossRef]
23. Wu, M.; Zhang, G.; Zhang, S.; Huang, W.; Zou, H. Spatiotemporal decoupling of rainfall patterns and shallow landslide stability: Lessons learned from calibrated ensemble physically-based models. *J. Rock Mech. Geotech. Eng.* **2025**. [CrossRef]
24. Xiao, S.; Xiao, T.; Jiang, R.; Wang, H.; Ju, L.; Zhang, L. Two-phase strategy for rapid and unbiased assessment of earthquake-induced landslides. *Eng. Geol.* **2024**, *336*, 107562. [CrossRef]
25. Liu, S.; Wang, L.; Zhang, W.; Sun, W.; Wang, Y.; Liu, J. Physics-informed optimization for a data-driven approach in landslide susceptibility evaluation. *J. Rock Mech. Geotech. Eng.* **2024**, *16*, 3192–3205. [CrossRef]
26. Mo, W.; Zhao, Y.; Yang, N.; Xu, Z.; Zhao, W.; Li, F. Effects of climate and land use/land cover changes on water yield services in the Dongjiang Lake Basin. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 466. [CrossRef]
27. Yang, J.; Dong, J.; Xiao, X.; Dai, J.; Wu, C.; Xia, J.; Zhao, G.; Zhao, M.; Li, Z.; Zhang, Y.; et al. Divergent shifts in peak photosynthesis timing of temperate and alpine grasslands in China. *Remote Sens. Environ.* **2019**, *233*, 111395. [CrossRef]
28. Wang, M.; Yao, J.; Chang, H.; Liu, R.; Cao, Y.; Zhao, Y. Monthly Groundwater Level Grid Dataset of China Region (2005-2022). Natl. Tibet. Plateau/Third Pole Environ. Data Cent. 2024. Available online: https://cstr.cn/18406.11.Terre.tpdc.301342 (accessed on 20 January 2025).
29. Leonarduzzi, E.; McArdell, B.W.; Molnar, P. Rainfall-induced shallow landslides and soil wetness: Comparison of physically-based and probabilistic predictions. *Hydrol. Earth Syst. Sci. Discuss.* **2020**, *25*, 5937–5950. [CrossRef]
30. Pradhan, B.; Lee, S. Delineation of landslide hazard areas on Penang Island, Malaysia, by using frequency ratio, logistic regression, and artificial neural network models. *Environ. Earth Sci.* **2010**, *60*, 1037–1054. [CrossRef]
31. Gardner, W.R. Some steady-state solutions of the unsaturated moisture flow equation with application to evaporation from a water table. *Soil Sci.* **1958**, *85*, 228–232. [CrossRef]
32. Srivastava, R.; Yeh, T.-C.J. Analytical solutions for one-dimensional, transient infiltration toward the water table in homogeneous and layered soils. *Water Resour. Res.* **1991**, *27*, 753–762. [CrossRef]
33. Yang, L.; Cui, Y.; Xu, C.; Ma, S. Application of coupling physics-based model TRIGRS with random forest in rainfall-induced landslide-susceptibility assessment. *Landslides* **2024**, *21*, 2179–2193. [CrossRef]
34. Weidner, L.; Oommen, T.; Escobar-Wolf, R.; Sajinkumar, K.S.; Samuel, R.A. Regional-scale back-analysis using TRIGRS: An approach to advance landslide hazard modeling and prediction in sparse data regions. *Landslides* **2018**, *15*, 2343–2356. [CrossRef]
35. Dagdelenler, G.; Nefeslioglu, H.A.; Gokceoglu, C. Modification of seed cell sampling strategy for landslide susceptibility mapping: An application from the Eastern part of the Gallipoli Peninsula (Canakkale, Turkey). *Bull. Eng. Geol. Environ.* **2016**, *75*, 575–590. [CrossRef]
36. Godt, J.W.; Baum, R.L.; Savage, W.Z.; Salciarini, D.; Schulz, W.H.; Harp, E.L. Transient deterministic shallow landslide modeling: Requirements for susceptibility and hazard assessments in a GIS framework. *Eng. Geol.* **2008**, *102*, 214–226. [CrossRef]
37. Wu, Y.; Lan, H.; Gao, X.; Li, L.; Meng, Y. A regional slope stability assessment model based on Bayesian theory. *J. Eng. Geol.* **2014**, *22*, 1227–1233. [CrossRef]

https://doi.org/10.3390/rs18030408

38. Tan, D.; Xu, X.; Wang, L.; Xu, J.; Shi, Q. Deformation evolution and failure mechanism of rainfall-induced granite residual soil landsliding event in Northern Guangdong, China. *Landslides* **2025**, *22*, 925–941. [CrossRef]
39. Chen, T.; Sanjou, M.; Hiraishi, T.; Xu, G. Failure mechanisms of carbonate rock landslides: Structure, karst generation and reservoir water fluctuations. *Geomorphology 2025*, 109975. [CrossRef]
40. Sahin, E.K. Assessing the predictive capability of ensemble tree methods for landslide susceptibility mapping using XGBoost, gradient boosting machine, and random forest. *SN Appl. Sci.* **2020**, *2*, 1308. [CrossRef]
41. Huang, F.; Cao, Z.; Guo, J.; Jiang, S.-H.; Li, S.; Guo, Z. Comparisons of heuristic, general statistical and machine learning models for landslide susceptibility prediction and mapping. *Catena* **2020**, *191*, 104580. [CrossRef]
42. Tien Bui, D.; Tuan, T.A.; Klempe, H.; Pradhan, B.; Revhaug, I. Spatial prediction models for shallow landslide hazards: A comparative assessment of the efficacy of support vector machines, artificial neural networks, kernel logistic regression, and logistic model tree. *Landslides* **2016**, *13*, 361–378. [CrossRef]
43. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
44. Zhao, F.; Miao, F.; Wu, Y.; Ke, C.; Gong, S.; Ding, Y. Refined landslide susceptibility mapping in township area using ensemble machine learning method under dataset replenishment strategy. *Gondwana Res.* **2024**, *131*, 20–37. [CrossRef]
45. Wu, R.Z.; Hu, X.D.; Mei, H.B.; He, J.; Yang, J. Spatial susceptibility assessment of landslides based on random forest: A case study from Hubei section in the Three Gorges Reservoir area. *Earth Sci.* **2021**, *46*, 321–330. [CrossRef]
46. Huang, F.; Tao, S.; Chang, Z.; Huang, J.; Fan, X.; Jiang, S.-H.; Li, W. Efficient and automatic extraction of slope units based on multi-scale segmentation method for landslide assessments. *Landslides* **2021**, *18*, 3715–3731. [CrossRef]
47. Xie, C.; Huang, Y.; Li, L.; Li, T.; Xu, C. Detailed inventory and spatial distribution analysis of rainfall-induced landslides in Jiexi County, Guangdong Province, China in August 2018. *Sustainability* **2023**, *15*, 13930. [CrossRef]
48. Kavzoglu, T.; Sahin, E.K.; Colkesen, I. Landslide susceptibility mapping using GIS-based multi-criteria decision analysis, support vector machines, and logistic regression. *Landslides* **2014**, *11*, 425–439. [CrossRef]
49. Li, M.; Wang, H.; Chen, J.; Zheng, K. Assessing landslide susceptibility based on the random forest model and multi-source heterogeneous data. *Ecol. Indic.* **2024**, *158*, 111600. [CrossRef]
50. Li, Y.; Hu, X.; Zhang, H.; Zheng, H.; Li, N. Displacement prediction and failure mechanism analysis of rainfall-induced colluvial landslides. *J. Hydrol.* **2025**, 133361. [CrossRef]
51. Xu, Q.; Zhao, B.; Dai, K.; Dong, X.; Li, W.; Zhu, X.; Yang, Y.; Xiao, X.; Wang, X.; Huang, J.; et al. Remote sensing for landslide investigations: A progress report from China. *Eng. Geol.* **2023**, *321*, 107156. [CrossRef]
52. Guo, Z.; Tian, B.; Zhu, Y.; He, J.; Zhang, T. How do the landslide and non-landslide sampling strategies impact landslide susceptibility assessment?—A catchment-scale case study from China. *J. Rock Mech. Geotech. Eng.* **2024**, *16*, 877–894. [CrossRef]
53. Li, C.; Feng, P.; Meng, J.; Catani, F.; Hellevang, H.; Tang, H.; Sun, X.; Huang, D. Physics-informed deep learning for revealing the evolutionary characteristics of landslides induced by rainfall process. *Geophys. Res. Lett.* **2025**, *52*, e2025GL117356. [CrossRef]
54. Zhang, J.; Ma, X.; Zhang, J.; Sun, D.; Zhou, X.; Mi, C.; Wen, H. Insights into geospatial heterogeneity of landslide susceptibility based on the SHAP-XGBoost model. *J. Environ. Manag.* **2023**, *332*, 117357. [CrossRef] [PubMed]
55. Huang, J.; Wen, H.; Zhou, X.; Xiao, J. Is there difference in landslide susceptibility model based on explainable artificial intelligence from the perspective of slope units with different scales? *Reliab. Eng. Syst. Saf.* **2025**, 111701. [CrossRef]
56. Xu, F.; Xu, Q.; Pu, C.; Wang, X.; Xu, P. Can different machine learning methods have consistent interpretations of DEM-based factors in shallow landslide susceptibility assessments? *J. Rock Mech. Geotech. Eng.* **2025**, *17*, 7864–7881. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

https://doi.org/10.3390/rs18030408

fire

MDPI

Article

Early Wildfire Smoke Detection with a Multi-Resolution Framework and Two-Stage Classification Pipeline

Gihwan Jung [1], Tae-Hyuk Ahn [1] and Byungseok Min [2,*]

[1] Department of Computer Science, Saint Louis University, Saint Louis, MO 63103, USA
[2] Department of Artificial Intelligence Data Science, Sejong University, Seoul 05006, Republic of Korea
* Correspondence: bmin@sejong.ac.kr

Abstract

Early wildfire smoke detection is critical for preventing small ignitions from escalating into large-scale fires, yet early-stage smoke plumes are often faint, low-contrast, and spatially small. When full-resolution frames are resized to satisfy fixed-input detector architectures and enable efficient batched GPU inference, these subtle cues are further diminished, leading to missed detections and unreliable scores near deployment thresholds. Existing remedies such as multi-scale inference, slicing/tiling, or super-resolution could improve sensitivity, but typically incur substantial overhead from multiple forward passes or added network components, limiting real-time use on resource-constrained platforms. To mitigate these challenges, we propose a composite multi-resolution detection framework that improves sensitivity to small smoke regions while maintaining single-pass inference. Motivated by the fact that most operational wildfire monitoring systems rely on Unmanned Aerial Vehicle (UAV) platforms and mountain-top Closed-Circuit Television (CCTV) systems surveillance, their wide-field imagery typically contains a large sky region above the horizon where early smoke is most likely to first become visible. Accordingly, crop placement is guided by a skyline prior that prioritizes this high-probability sky band while retaining the remaining scene for global context. A dynamic compositing stage stacks a global view with a high-resolution, sky-aligned band into a standard square detector input, preserving context with minimal added cost. Detections from the two views are reconciled via coordinate restoration and non-maximum suppression. For deployment, a lightweight second-stage classifier selectively re-evaluates low-confidence detections to stabilize decisions near a fixed operating threshold without retraining the detector. Compared to the baseline detector, our approach improves detection performance on the Early Smoke dataset, achieving gains of +4.6 percentage points in AP @0.5:0.95, +3.4 percentage points in AP @0.5, +2.9 percentage points in precision, +5.3 percentage points in recall, and +4.3 percentage points in F1-score.

Keywords: early smoke detection; wildfire monitoring; small object detection; multiresolution

1. Introduction

Early fire detection plays a crucial role in preventing small ignition events from escalating into large-scale wildfires that cause extensive ecological damage, economic loss, and threats to human life and infrastructure [1]. Once a fire grows beyond its incipient stage, suppression becomes significantly more difficult and costly, particularly in remote or mountainous regions where access is limited and response times are long [2]. As climate change and prolonged drought conditions increase both the frequency and intensity of

wildfires, the ability to identify fire-related signals at the earliest possible moment has become a central requirement for modern wildfire management systems [3]. In operational settings, many detection systems rely on vision-based monitoring using unmanned aerial vehicles (UAVs) equipped with high-resolution red, green, and blue (RGB) cameras or fixed mountain-top closed-circuit television (CCTV) surveillance systems [4]. These platforms enable wide-area coverage and continuous observation without the need for dense ground sensor deployment, making single-image smoke detection a common and scalable paradigm [5]. However, early-stage smoke presents a uniquely challenging visual target: it often occupies only a very small portion of the image-a small fraction of the total area-while exhibiting low contrast, diffuse boundaries, and strong visual similarity to clouds, haze, or terrain [6]. Despite these difficulties, smoke is typically the first visible indicator of ignition, appearing well before flames are detectable, which makes robust early smoke detection a critical component of effective wildfire prevention and response [7].

With the rise of deep learning, vision-based object detectors became a central component of wildfire monitoring pipelines in both fixed-camera and aerial surveillance settings [8]. Convolutional neural network (CNN)-based architectures, including Region-based Convolutional Neural Networks (R-CNN) and You Only Look Once (YOLO) families, were widely deployed due to their favorable accuracy-efficiency characteristics, while more recent transformer-based detectors and hybrid CNN-transformer models further advanced robustness and contextual modeling across diverse vision tasks [9–11]. Building on these developments, contemporary wildfire detection systems achieved substantial progress under a range of operational conditions [12]. However, detecting early-stage smoke remains a persistent challenge, as nascent plumes are often faint, low-contrast, and spatially small, with diffuse boundaries that can closely resemble confounding visual phenomena such as clouds, haze, or terrain textures [13]. A practical consideration in many deployments is that high-resolution frames are resized to fixed input dimensions to satisfy architectural constraints and enable efficient batched inference, which can attenuate subtle smoke cues and reduce their salience in downstream feature representations [14]. To mitigate these effects, a broad set of resolution-aware and small-object detection strategies has been investigated, including multi-scale inference, high-resolution processing, and image slicing/tiling, and these approaches demonstrated measurable gains for small targets in related remote-sensing and surveillance contexts [15]. Nonetheless, because real-world wildfire monitoring systems must often operate under strict latency and resource constraints, further improvements remain desirable—particularly in the early-smoke regime-toward enhancing recall and stabilizing confidence behavior near deployment thresholds while maintaining computational practicality [12].

To address the limitations of existing wildfire detection systems in the early-smoke regime, we propose a cost-efficient composite detection framework that improves sensitivity to small, faint smoke while preserving single-pass inference. The approach reallocates spatial resolution by embedding a targeted high-resolution region within a standard detector input, guided by a skyline-based spatial prior to emphasize likely smoke regions while retaining global scene context. For deployment, we incorporate a lightweight confidence refinement step that operates only in a low-confidence range to improve prediction stability without changing training or standard evaluation. The proposed pipeline avoids expensive multi-stage slicing or multi-scale inference while remaining compatible with real-time, resource-constrained settings.

Our main contributions are:

- A composite multi-resolution representation that embeds high-resolution detail within a standard detector input using a single forward pass.

https://doi.org/10.3390/fire9020092

- A skyline-guided dynamic cropping strategy with reversible coordinate mapping for accurate localization in the original image space.
- A deployment-time confidence refinement mechanism to stabilize low-confidence early-smoke predictions.
- A comprehensive evaluation on early-smoke benchmarks (Early Smoke and Pyro-SDIS) demonstrating consistent gains in small-object and low-confidence regimes.

The remainder of this paper is organized as follows. The Section 2 reviews prior research in wildfire monitoring, deep learning for wildfire detection, and small-object detection with resolution manipulation. The Section 3 describes the proposed multi-resolution detection framework and the two-stage confidence refinement pipeline. The Section 4 details the datasets, training settings, and evaluation metrics. The Section 5 reports experimental results, including deployment-only evaluation, ablation studies, and generalizability verification. Finally, Section 6 summarizes the main findings and implications for early-smoke detection.

2. Related Work

2.1. Wildfire Monitoring Context

Wildfires arise from a combination of ignition sources and conducive environmental conditions, including low fuel moisture, high winds, and complex topography. Ignitions may be natural (e.g., lightning) or human-caused (e.g., equipment use, power infrastructure, and debris burning), and even small ignitions can rapidly evolve into large incidents under adverse weather and fuel conditions [16]. Once established, wildfires can spread quickly and produce severe societal and ecological impacts, including loss of life and property, disruption to critical infrastructure, long-term ecosystem damage, and widespread degradation of air quality due to smoke transport [17]. These risks motivate an operational emphasis on early warning, where detecting nascent smoke plumes can enable faster dispatch and containment before initial ignitions escalate. Operational wildfire management relies on a combination of prevention, preparedness, and rapid response, supported by monitoring and situational awareness across multiple sensing modalities. Existing monitoring approaches are broadly categorized into human-based and autonomous systems. Human-based monitoring includes lookout towers, patrol flights, and public or field personnel reporting, which remain valuable but are labor-intensive, costly to scale, and inherently limited in temporal continuity and spatial coverage [18]. These limitations motivate increasing interest in automated alternatives that can operate persistently.

Autonomous monitoring systems span a range of sensing platforms operating at different spatial and temporal resolutions. Satellite-based systems provide wide-area coverage and routine global observation, making them effective for regional situational awareness but often limited in spatial resolution and revisit frequency for early-stage detection [19]. Aerial platforms, including crewed aircraft and unmanned areial vehicles (UAVs), provide flexible, high-resolution sensing and rapid deployment capabilities, while their operational characteristics naturally introduce considerations related to cost, endurance, and weather dependence [20]. Ground-based vision systems, such as fixed CCTV camera networks deployed on towers or mountain ridges, enable continuous local surveillance and real-time visual confirmation, making them well suited for persistent monitoring of high-risk regions, while their effective coverage is influenced by terrain and camera placement [21]. Across these autonomous platforms, a wide range of analytical tools was explored. Early rule-based approaches relied on handcrafted color and motion heuristics to detect fire or smoke in video sequences [22]. Subsequent statistical learning approaches replaced fixed rules with engineered features and trained classifiers; for example, wavelet-based smoke characterization combined with support vector machines demonstrated reliable

https://doi.org/10.3390/fire9020092

early detection with low false-alarm rates across surveillance scenarios [23]. More recently, artificial intelligence (AI)-driven surveillance systems based on machine learning and deep learning gained growing attention, enabling end-to-end automated detection, reducing human workload and improving the timeliness and robustness of wildfire monitoring across heterogeneous sensing modalities [24].

2.2. Deep Learning for Wildfire Monitoring

Recent deep learning-based wildfire monitoring research was organized by the primary vision task and system structure employed [25]. A first line of work formulated wildfire monitoring as an image-level classification problem, predicting whether a frame contains fire or smoke to enable rapid triage under limited computation. For example, Akagic et al. [26] proposed LW-FIRE, a lightweight CNN designed for real-time wildfire image classification with dataset transformation to improve generalization. Seydi et al. [27] introduced Fire-Net, a deep learning–based remote sensing framework that integrates optical and thermal information to enable accurate and robust detection of active forest fires across diverse environments, including small-scale fire events.

A second line focused on object detection, aiming to localize fire or smoke regions with bounding boxes for spatially explicit situational awareness and downstream tracking [28]. MModern vision systems for wildfire monitoring are commonly built on established object-detection families such as R-CNN and YOLO [9–11], which offered a practical balance between accuracy and real-time speed, alongside newer transformer-based and hybrid CNN-transformer designs that could better leverage broader scene context. Li et al. [29] proposed LEF-YOLO, a lightweight detector tailored for real-time identification of extreme wildfire cases and shown to perform effectively on a specialized dataset while maintaining fast inference. Bhargav and Singh [30] proposed a UAV-based wildfire detection framework that combines a CNN-based image classification model with YOLOv8 for object localization, demonstrating improved accuracy and computational efficiency over traditional methods while enabling timely fire detection from UAV imagery under challenging visual conditions.

A third line adopted semantic segmentation, producing pixel-level masks that could better represent irregular boundaries and diffuse phenomena, particularly for smoke plumes or elongated flame fronts. Jonnalagadda and Hashim [31] developed a segmentation-based UAV pipeline optimized for real-time flame detection. Khryashchev and Larionov [32] developed a satellite-image wildfire segmentation model based on U-Net with a ResNet34 encoder, trained with augmentation on high-resolution RGB imagery and evaluated using Dice and intersection over union (IoU) metrics to support early wildland fire detection.

Beyond single-task formulations, many practical systems used cascaded or multi-stage networks [33] that compose these tasks-most commonly using a lightweight classifier or coarse detector to propose candidate regions followed by a higher-capacity detector or segmenter to refine localization-improving robustness while controlling inference cost across high-resolution imagery. Ghali et al. [34] proposed a UAV-based wildfire framework that combines a deep ensemble classifier built on EfficientNet-B5 and DenseNet-201 with segmentation architectures including TransUNet, TransFire, and EfficientSeg. The framework enables accurate wildfire classification and precise pixel-level delineation, remaining robust to small fire regions and complex aerial backgrounds.

Collectively, these approaches have substantially improved the reliability of fire and smoke recognition in operational imagery; however, there remains significant scope for further improvement. In particular, a large portion of prior work is evaluated on scenarios in which the target signal is visually salient (e.g., large flames, dense smoke plumes, or fully developed fire fronts), where recognition cues are strong and localization is comparatively

https://doi.org/10.3390/fire9020092

less ambiguous. Consequently, reported gains in classification accuracy, detection mean average precision (mAP), or segmentation overlap may not fully characterize performance in settings that are most critical for proactive monitoring, where the evidence is weaker and more readily confounded by background structure and atmospheric phenomena. Moreover, common efficiency-driven preprocessing practices—most notably resizing high-resolution frames to fixed input resolutions—may suppress fine-grained spatial detail that is informative for subtle visual signatures, motivating continued research on methods that generalize beyond prominent fire imagery while remaining computationally practical for real-time deployment.

2.3. Small Object Detection and Resolution Manipulation

Broadly, early wildfire detection shares key similarities with small object detection because the first visible cues-nascent smoke wisps or small flames-often occupy very few pixels and exhibit weak visual evidence. These cues are further degraded by downsampling and compression and are often confused with clouds, fog, dust, sun glare, or terrain textures, making them difficult to detect reliably in cluttered or low-contrast scenes. Remote-sensing surveys such as Cheng et al. [35] highlighted that even strong multi-scale and context-aware designs could struggle when targets became extremely small, which aligns closely with the "tiny smoke" regime encountered in early warning settings.

Consequently, detection systems for small objects tended to intervene on scale in two complementary ways. First, some works strengthened the detector itself to be more responsive to small targets, often through improved multi-scale handling and better localization behavior in the tiny-object regime [36–39]. Second, many practical pipelines manipulated the input or processing strategy so that early smoke or flame signals remained visible to the model without requiring uniformly higher resolution over the entire frame. Tiling and slicing approaches [40,41] and their refinements [42–46] reflected a common operational insight: allocating more spatial detail to candidate regions could materially improve detection when the event occupied only a small portion of the scene. Related region-adaptive methods similarly prioritized resolution where it mattered most by focusing computation on predicted areas of interest rather than treating all pixels equally [47].

Collectively, these directions improved small-target performance in fire monitoring, but they also exposed trade-offs that matter for early warning. Resolution-based pipelines may introduce additional runtime and engineering complexity due to repeated per-region inference and merging, and their performance may depend on design choices such as tile size and overlap. Meanwhile, detector-only improvements may raise sensitivity broadly across the full frame rather than explicitly preserving fine detail only in likely smoke regions. These issues are especially pronounced for early smoke, which is diffuse and low-contrast and may occupy only a handful of pixels, motivating wildfire-specific strategies that allocate resolution and computation selectively while maintaining real-time practicality.

Our Approach

In summary, prior wildfire detection systems have attempted to improve early-smoke performance mainly through (i) detector-centric architectural improvements for small objects, (ii) resolution-heavy strategies such as multi-scale inference or tiling/slicing, and (iii) multi-stage pipelines that cascade classification, detection, or segmentation. While these approaches can raise sensitivity, they often increase runtime and system complexity, and their effectiveness can be limited in the early-smoke regime where cues are faint, diffuse, and easily confounded—especially after high-resolution frames are resized to fixed detector inputs.

https://doi.org/10.3390/fire9020092

In contrast, the proposed framework targets this gap by selectively reallocating spatial resolution rather than uniformly increasing computation. It embeds a targeted high-resolution region within a standard detector input, guided by a skyline-based spatial prior, preserving fine detail in likely smoke regions while maintaining global context and single-pass inference (avoiding multi-tile or multi-scale passes). In addition, instead of adding training-time stages, it introduces a lightweight deployment-time confidence refinement applied only in a low-confidence range to stabilize borderline early-smoke predictions without changing training or standard evaluation, improving practicality for real-time, resource-constrained monitoring.

3. Methods

3.1. Overall Architecture

Figure 1 provides an overview of the proposed multi-resolution detection framework. The overall system is composed of three main components: (1) composite image generation with a sky-adaptive logic module for skyline-based cropping, (2) object detection on the composite input, and (3) a fusion stage that performs coordinate remapping, non-maximum suppression (NMS), and optionally a two-stage classifier for deployment. The following subsections describe each component in detail.

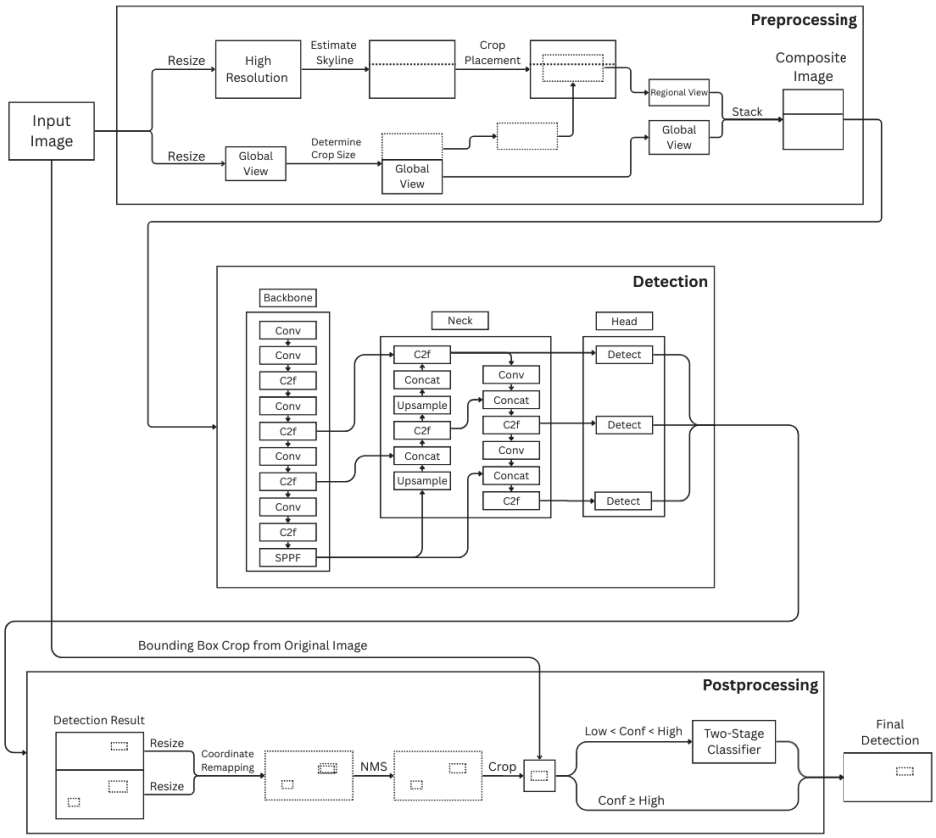


Figure 1. Overview of the proposed multi-resolution detection framework.

3.2. Composite Image Generation

Given an input image I, most modern object detectors operate on a fixed-size square input of $W_{in} \times H_{in}$ to enable efficient batching and to maintain a consistent feature-map geometry throughout the network. When the aspect ratio of I differs from the required square shape, the common practice is to resize I while preserving its aspect ratio and then apply letterboxing, i.e., padding the remaining area (typically with constant-valued gray

https://doi.org/10.3390/fire9020092

or black pixels) so that the final tensor matches the $W_{in} \times H_{in}$ input constraint. These padded regions contribute negligible semantic information and primarily serve as a shape-normalization mechanism since they contain no scene content.

In this study, the otherwise uninformative padding area is repurposed as additional representational capacity. Instead of leaving this region as padding, the square canvas is filled with an auxiliary high-detail view of the same scene. This design preserves a global-context image while allocating more pixels to regions expected to contain early smoke signatures. The approach increases the effective spatial resolution available for small and visually ambiguous targets without changing the detector input size, modifying the backbone architecture, or incurring the computational overhead associated with multi-pass tiling or slicing-based inference (see Figure 2).

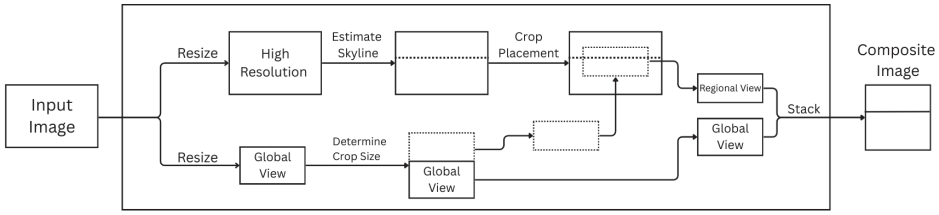


Figure 2. Composite image generation pipeline diagram.

To construct the composite representation, the original image I is first resized to a width of W_{in} pixels using bilinear interpolation while preserving aspect ratio, producing an image I_{global} of size $W_{in} \times H_{\text{global}}$. A second resized version I_{inter} is generated from the same high-resolution source using bilinear interpolation at a larger intermediate width S_{int}, allowing the system to retain object-level details at a higher spatial resolution. These two resized images provide the global and high-resolution views used in the composite. The cropping window is set to match the unused vertical space after resizing, i.e., a window of size $W_{in} \times (H_{in} - H_{\text{global}})$, so that stacking it with I_{global} yields a final $W_{in} \times H_{in}$ square. The vertical placement is determined using the estimated skyline position S, ensuring that the crop aligns with the region where smoke typically appears, and the crop is horizontally centered so that its midpoint aligns with the center of I_{inter}. Once the crop I_{roi} is extracted, it is placed above I_{global} to form the final composite image I_{comp}. The result is a $W_{in} \times H_{in}$ image composed of a top band containing a high-resolution regional view derived from I_{inter} and a bottom band containing the global view.

During composite construction, a set of metadata parameters required for mapping detection boxes from composite coordinates back to the original image space is stored. These quantities capture the intermediate resizing scales, cropping boundaries, and any padding introduced during stacking. Let

- H_{gap} denotes the crop height,
- (m_0, n_0) the horizontal and vertical crop offsets,
- s_{int} and $s_{W_{in}}$ the scaling factors used for generating the intermediate-resolution and W_{in}-width images,
- $(S_{\text{int}}, H_{\text{inter}})$ the dimensions of the intermediate resized image,
- o_{top} and o_{bot} the vertical offsets introduced by cropping when mapping detections back to the original coordinate space.

These metadata variables are passed to the fusion stage, where they enable precise restoration of bounding boxes to the coordinate system of the original image I. The overall procedure is summarized in Algorithm 1.

https://doi.org/10.3390/fire9020092

Algorithm 1 Composite Image Generation Procedure.

Procedure GenerateComposite$(I, W_{in}, H_{in}, S_{\text{int}})$:

Input: I (RGB image), W_{in}, H_{in} (detector input size), S_{int} (intermediate resize width).
Output: $I_{\text{comp}} \in \mathbb{R}^{W_{in} \times H_{in} \times 3}$ (composite for detector inference),
M (metadata for coordinate restoration).

1. **Global view (base canvas for detector input):**
$I_{\text{global}} \leftarrow \text{ResizeWidth}(I, W_{in})$ (preserve aspect ratio),
$I_{\text{global}} \in \mathbb{R}^{W_{in} \times H_{\text{global}} \times 3}$.

2. **Intermediate view (used to select a smoke-prior ROI):**
$I_{\text{inter}} \leftarrow \text{ResizeWidth}(I, S_{\text{int}})$ (preserve aspect ratio),
$I_{\text{inter}} \in \mathbb{R}^{S_{\text{int}} \times H_{\text{inter}} \times 3}$.

3. **Define the ROI size to occupy the unused letterbox region:**
$H_{\text{gap}} \leftarrow H_{in} - H_{\text{global}}$,
$(w_{\text{roi}}, h_{\text{roi}}) \leftarrow (W_{in},\ H_{\text{gap}})$.

4. **Compute ROI location (skyline-guided, with fallback):**
$S \leftarrow \text{SkylineRow}(I_{\text{inter}})$.
if S is undefined: $S \leftarrow \lfloor \eta H_{\text{inter}} \rfloor \quad (\eta \in (0,1)$ is a default prior$)$.
$m_0 \leftarrow \left\lfloor \frac{S_{\text{int}} - w_{\text{roi}}}{2} \right\rfloor, \quad m_1 \leftarrow m_0 + w_{\text{roi}}$.
$n_0 \leftarrow \text{clip}(\lfloor S - \beta h_{\text{roi}} \rfloor,\ 0,\ H_{\text{inter}} - h_{\text{roi}})$,
$n_1 \leftarrow n_0 + h_{\text{roi}}$.
Define $\text{ROI}_{\text{inter}} \leftarrow [m_0, m_1) \times [n_0, n_1)$.

5. **Crop the intermediate view:**
$I_{\text{roi}} \leftarrow \text{Crop}(I_{\text{inter}},\ \text{ROI}_{\text{inter}})$,
$I_{\text{roi}} \in \mathbb{R}^{W_{in} \times H_{\text{gap}} \times 3}$.

6. **Construct the detector input (single $W_{in} \times H_{in}$ tensor):**
$I_{\text{comp}} \leftarrow \text{Concat}_{\text{vertical}}(I_{\text{roi}},\ I_{\text{global}})$,
$I_{\text{comp}} \in \mathbb{R}^{W_{in} \times H_{in} \times 3}$.

7. **Return composite and restoration metadata:**
$M \leftarrow \left\{ W_{in},\ H_{in},\ S_{\text{int}},\ H_{\text{global}},\ H_{\text{gap}},\ (m_0, n_0),\ \text{resize scales for } I_{\text{global}} \text{ and } I_{\text{inter}} \right\}$.
return (I_{comp}, M).

3.3. Dynamic Cropping

To obtain a stable reference point for placing the high-resolution crop, the skyline is estimated by leveraging color statistics in the YCrCb color space. The purpose of this module is to avoid placing the crop window without guidance and instead use a consistent structural feature of outdoor scenes. The sky generally occupies the upper portion of the image and exhibits distinctive luminance–chrominance values; therefore, the transition between sky and terrain serves as a practical reference for positioning the higher-resolution crop. Figure 3 illustrates the intermediate and final outputs of this procedure.

To identify the skyline region, the high-resolution input image I_{inter} is first downsampled by a factor $\alpha \in (0, 1)$ to reduce computational cost, producing $I_{\text{inter},\alpha}$ whose width and height are scaled by α. The downsampled image $I_{\text{inter},\alpha}$ is then converted from RGB to YCrCb. In practice, we set $\alpha \in [0.125, 0.5]$ based on empirical validation, which substantially reduces computation while preserving accurate skyline estimation.

https://doi.org/10.3390/fire9020092

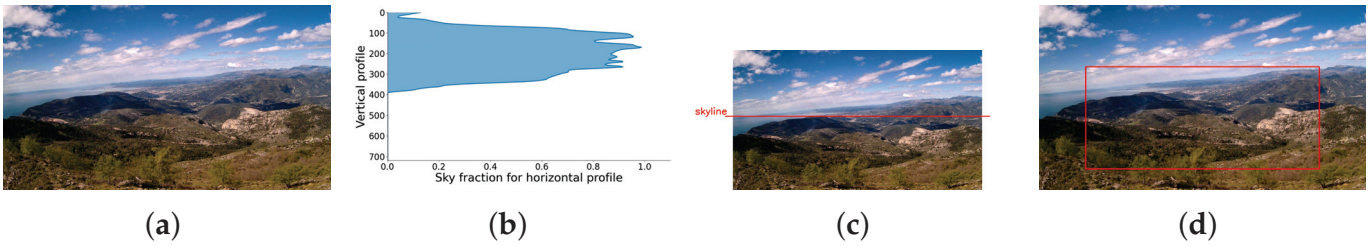


Figure 3. Step-by-step visualization of skyline detection. The module converts the image to YCrCb, identifies sky pixels via chrominance–luminance thresholds, and detects the sky-to-land transition through row-wise analysis. (**a**) Input image, (**b**) Sky pixel distribution, (**c**) Skyline detection, (**d**) ROI extraction.

A binary mask is subsequently constructed by labeling pixels as "sky" according to luminance and chrominance criteria. Specifically, a pixel is marked as sky if its luminance exceeds the global mean of the Y channel and its chrominance values fall within empirically determined thresholds. The RGB-to-YCrCb conversion is calculated as:

$$\begin{aligned} Y &= 0.299R + 0.587G + 0.114B, \\ Cb &= 128 - 0.168736R - 0.331264G + 0.5B, \\ Cr &= 128 + 0.5R - 0.418688G - 0.081312B, \end{aligned}$$

where $R, G, B \in [0, 255]$.

To determine the Cr and Cb thresholds, we assume $Y \in [0, 255]$ and partition the luminance range into three intervals using four breakpoints, $Y_1 = 64$, $Y_2 = 128$, $Y_3 = 160$, and $Y_4 = 192$. Sky regions are typically associated with higher luminance; therefore, pixels with $Y < Y_1$ (i.e., $Y < 64$) are discarded and are not considered as sky candidates. At each breakpoint Y_k, define empirically tuned chrominance thresholds

$$\Theta^{(k)} = \left(\tau_{\mathrm{Cb}}^{\min,(k)}, \tau_{\mathrm{Cb}}^{\max,(k)}, \tau_{\mathrm{Cr}}^{\min,(k)}, \tau_{\mathrm{Cr}}^{\max,(k)}\right), \qquad k = 1, 2, 3, 4.$$

Each breakpoint threshold set $\Theta^{(k)}$ is initialized from dataset statistics by taking the sample mean and standard deviation of the sky-pixel chrominance distributions at luminance level Y_k, and setting each bound to lie within two standard deviations of the mean (i.e., $\mu \pm 2\sigma$). The final values are then selected empirically based on validation performance. For a pixel luminance $Y_{m,n}$, let $k \in \{1, 2, 3\}$ satisfy $Y_{m,n} \in [Y_k, Y_{k+1}]$, and define

$$\alpha_k(Y_{m,n}) = \frac{Y_{m,n} - Y_k}{Y_{k+1} - Y_k} \in [0, 1].$$

Then the luminance-adaptive thresholds are given by piecewise linear interpolation:

$$\begin{aligned} \tau_{\mathrm{Cb}}^{\min}(Y_{m,n}) &= (1 - \alpha_k)\, \tau_{\mathrm{Cb}}^{\min,(k)} + \alpha_k\, \tau_{\mathrm{Cb}}^{\min,(k+1)}, \\ \tau_{\mathrm{Cb}}^{\max}(Y_{m,n}) &= (1 - \alpha_k)\, \tau_{\mathrm{Cb}}^{\max,(k)} + \alpha_k\, \tau_{\mathrm{Cb}}^{\max,(k+1)}, \\ \tau_{\mathrm{Cr}}^{\min}(Y_{m,n}) &= (1 - \alpha_k)\, \tau_{\mathrm{Cr}}^{\min,(k)} + \alpha_k\, \tau_{\mathrm{Cr}}^{\min,(k+1)}, \\ \tau_{\mathrm{Cr}}^{\max}(Y_{m,n}) &= (1 - \alpha_k)\, \tau_{\mathrm{Cr}}^{\max,(k)} + \alpha_k\, \tau_{\mathrm{Cr}}^{\max,(k+1)}. \end{aligned}$$

The binary sky mask M is defined as:

$$\begin{aligned} M_{m,n} = \mathbb{I}\big[\, Y_{m,n} > \mu_Y \;\wedge\; \tau_{\mathrm{Cr}}^{\min}(Y_{m,n}) \le \mathrm{Cr}_{m,n} \le \tau_{\mathrm{Cr}}^{\max}(Y_{m,n}) \\ \wedge\; \tau_{\mathrm{Cb}}^{\min}(Y_{m,n}) \le \mathrm{Cb}_{m,n} \le \tau_{\mathrm{Cb}}^{\max}(Y_{m,n}) \,\big], \end{aligned}$$

https://doi.org/10.3390/fire9020092

where μ_Y denotes the average luminance across the image.

The row-wise profile is formalized by letting $r \in \{0, 1, \ldots, H_{\text{inter},\alpha} - 1\}$ denote the image row index (measured from the top). Let $n_{\text{sky}}(r)$ be the number of pixels labeled as sky in row r. The row-to-row change in sky-pixel counts is then computed as $\Delta n_{\text{sky}}(r) = n_{\text{sky}}(r+1) - n_{\text{sky}}(r)$. This row-wise sky-pixel profile $n_{\text{sky}}(r)$, together with its discrete difference $\Delta n_{\text{sky}}(r)$, forms the basis for detecting abrupt transitions used to estimate the skyline location.

To verify that this transition is meaningful, we compare the total sky-pixel mass above and below a candidate skyline row S using the row-wise sky counts $n_{\text{sky}}(r)$. Specifically, we compute $S_{\text{above}} = \sum_{r<S} n_{\text{sky}}(r)$ and $S_{\text{below}} = \sum_{r>S} n_{\text{sky}}(r)$. If $S_{\text{above}}/S_{\text{below}} > \text{Th}_{sky}$ for a threshold Th_{sky}, the candidate S is accepted as the skyline; otherwise, no skyline is returned. The estimated skyline position is then used to guide the vertical placement of the cropping window in the composite generation process.

Given a detected sky boundary at S, the crop window with height H_{gap} is placed such that $n_{\text{top}} = S - H_{\text{gap}}/4$, ensuring that only a small portion of the crop window contains sky and the majority covers land or forest regions where smoke is more likely to originate. If no valid skyline is detected, the crop window is instead centered vertically within the image to provide a neutral fallback that avoids systematic bias.

3.4. Fusion: Remapping and Non-Maximum Suppression

Each detection need to be mapped back to the coordinate system of the original image, as inference is performed on the $W_{in} \times H_{in}$ composite image. This remapping reverses the sequence of resizing, cropping, and padding operations applied during composite construction. For every predicted bounding box, the appropriate transformation branch is selected based on whether the detection originates from the top band (high-resolution crop) or the bottom band (global view). The metadata recorded during composite generation provides the offsets, crop coordinates, and scaling factors required for restoration.

$(\hat{m}_1, \hat{n}_1, \hat{m}_2, \hat{n}_2)$ is used to denote the bounding box coordinates in composite-image pixels, with $\hat{m}_1$ and $\hat{m}_2$ corresponding to the left and right column indices and $\hat{n}_1$ and $\hat{n}_2$ corresponding to the top and bottom row indices. The corresponding coordinates in the original image are computed by $m_k^{\text{orig}} = \frac{\hat{m}_k - o_m + c_m}{s_m}$, $n_k^{\text{orig}} = \frac{\hat{n}_k - o_n + c_n}{s_n}$, with $k \in \{1, 2\}$. where (o_m, o_n) represent the padding offsets applied to the composite band, (c_m, c_n) represent the crop offsets in the intermediate-resolution image, and s_m, s_n denote the scaling factors used to produce the resized inputs. For detections originating from the top band (the high-resolution crop), the remapping parameters are $o_m = o_{\text{top}}$, $o_n = 0$, $c_m = m_0$, $c_n = n_0$, and $s_m = s_n = s_{\text{int}}$. For detections originating from the bottom band (the globally resized image), the corresponding parameters are $o_m = o_{\text{bot}}$, $o_n = H_{\text{gap}}$, $c_m = 0$, $c_n = 0$, and $s_m = s_n = s_{W_{in}}$. This ensures that all geometric transformations applied during composite construction are correctly reversed, allowing detections from both the high-resolution and global views to be restored to their precise locations in the coordinate frame of the original image I.

After remapping detections from both the top and bottom bands to the original image coordinates, duplicate bounding boxes may occur because the composite pipeline effectively simulates two views of the same scene. To consolidate these detections, standard NMS is applied with an IoU threshold τ_{NMS}. Let b_m denote the m-th bounding box, s_m its confidence score, and $\text{IoU}(b_m, b_n)$ the intersection-over-union between boxes b_m and b_n. Let N denote the total number of boxes. The IoU is defined as

$$\text{IoU}(b_m, b_n) = \frac{|b_m \cap b_n|}{|b_m \cup b_n|}.$$

https://doi.org/10.3390/fire9020092

Given an IoU threshold $\tau \in (0, 1)$, the set of selected boxes after NMS is

$$S = \{m \in \{1, \ldots, N\} : \forall n \in \{1, \ldots, N\},\ s_n > s_m \Rightarrow \mathrm{IoU}(b_m, b_n) < \tau_{\mathrm{NMS}}\}.$$

The subset S forms the final set of detections for each image, expressed in the original coordinate frame.

3.5. Two-Stage Classifier

In deployment settings, a fixed confidence threshold is typically required to suppress false alarms. However, early-stage smoke plumes are typically faint, weakly textured, and visually diffuse, often occupying only a small portion of the image, which leads the detector to produce scores that cluster near the selected confidence threshold. Relying on a single cutoff therefore creates an inherent trade-off: raising the threshold improves precision but suppresses many true smoke detections, while lowering it increases recall at the cost of additional false positives. To alleviate this limitation, a lightweight two-stage classifier is introduced to refine only those detections whose confidence lies in an "uncertainty band" where YOLO predictions tend to be unreliable.

Detections with confidence below Th_{low} are treated as noise and discarded, whereas detections above Th_{high} are accepted without further processing. Only detections falling within the intermediate interval $[Th_{\mathrm{low}}, Th_{\mathrm{high}}]$ are passed to the classifier. This selective refinement follows the intuition of cascade-style detection frameworks, where ambiguous proposals receive additional evaluation, but differs in that our classifier is applied solely during deployment and does not modify detector training. For each detection requiring refinement, a square crop of size $r_{\mathrm{cls}} \times r_{\mathrm{cls}}$ is extracted from the original image such that the bounding-box center (m_c, n_c) lies at the center of the crop. Centering ensures that the classifier receives the local smoke structure along with the immediate surrounding context, while using a fixed resolution provides stable appearance statistics regardless of the underlying smoke size. The classifier then predicts whether the crop corresponds to foreground (smoke) or background.

The final decision rule combines the detector confidence with the classifier output. Detections with $\mathrm{conf}_{\mathrm{det}} < Th_{\mathrm{low}}$ are discarded outright, and detections with $\mathrm{conf}_{\mathrm{det}} > Th_{\mathrm{high}}$ are accepted immediately. For detections in the intermediate band, the classifier determines whether they are retained or removed. Computational overhead is minimized by limiting this refinement to a small subset of detections, so the additional computation is incurred only sporadically rather than uniformly across frames. Consequently, the pipeline maintains real-time throughput while improving robustness in cases where the detector alone yields insufficiently reliable predictions.

$$\mathrm{Decide} = \begin{cases} \text{DISCARD}, & \mathrm{conf}_{\mathrm{det}} < Th_{\mathrm{low}}, \\ \text{KEEP}, & \mathrm{conf}_{\mathrm{det}} > Th_{\mathrm{high}}, \\ \text{KEEP}, & Th_{\mathrm{low}} \le \mathrm{conf}_{\mathrm{det}} \le Th_{\mathrm{high}} \text{ and } \mathrm{cls}(d) = \mathrm{fg}, \\ \text{DISCARD}, & \text{otherwise.} \end{cases}$$

4. Experimental Setup

4.1. Datasets

4.1.1. Dataset Composition

Publicly available wildfire datasets contain relatively few instances of early-stage smoke, and the majority of images depict large, fully developed fires that are not representative of early-detection conditions. Early-smoke scenes are inherently rare in real-world collections because ignition events are seldom captured precisely at onset, and smoke

https://doi.org/10.3390/fire9020092

plumes remain visually faint for only a brief temporal window. In addition, many datasets include large numbers of near-duplicate frames, which can artificially inflate evaluation metrics and obscure true detector performance, particularly for small, low-contrast objects. Even when early-smoke images are present, the number of such samples is typically insufficient to support effective training. To address these limitations, a curated Early Smoke dataset is constructed from the FASDD-UAV and D-Fire collections. The Pyro-SDIS dataset is also employed, after deduplication, as a secondary benchmark to assess cross-dataset reliability.

4.1.2. Datasets Used

FASDD–UAV. FASDD is a large-scale flame and smoke detection dataset composed of three subsets (CV, UAV, RS) collected from terrestrial and airborne sensors such as surveillance cameras, watchtowers, and UAVs [48]. The UAV subset, FASDD–UAV, is used, which contains 25,097 UAV-captured images. Each image is annotated with bounding boxes in four formats (YOLO, PASCAL Visual Object Classes (VOC), Microsoft Common Objects in Context (COCO), and Training Data Markup Language (TDML)) for two object classes, *fire* and *smoke*, totaling 36,308 fire instances and 17,222 smoke instances. Image aspect ratios are primarily between 4:3 and 16:9. The dataset offers high-resolution aerial coverage but includes many large plumes and near-duplicate video frames; as a result, early-smoke examples are relatively scarce.

D-Fire. D-Fire is an image dataset for fire and smoke detection, comprising 21,527 images with 26,557 annotated bounding boxes, of which 11,865 are labeled as smoke and 14,692 as fire [49]. The classes are intentionally broad: smoke can appear concentrated or highly diffused, with color and intensity varying according to wind conditions and the materials being burned. Images were collected from the Internet, controlled fire simulations in the Technological Park of Belo Horizonte, surveillance cameras at the Universidade Federal de Minas Gerais (UFMG), and cameras at the Serra Verde State Park in Belo Horizonte. Additional synthetic images were created by compositing smoke patches onto green landscape backgrounds, a form of augmentation previously shown to be effective for training on synthetic smoke images. Annotations follow the YOLO format with normalized coordinates in $[0, 1]$, and the dataset spans diverse fire and smoke patterns, scenarios, camera viewpoints, and lighting conditions. As with `FASDD_UAV`, many scenes contain large flames; while some high-resolution early-smoke examples exist, most images depict well-developed fires, thus additional filtering is necessary for early-smoke-focused experiments.

Pyro-SDIS. Pyro-SDIS is a wildfire smoke detection dataset developed with the French Fire and Rescue Services (SDIS) and volunteers from the Pyronear association [50]. Images are captured from fixed detection towers equipped with 4–5 high-resolution cameras and an on-site microcomputer, monitoring mostly forested environments in wildfire-prone areas. Smoke is typically far from the camera, small in the image, and often low-contrast under diverse lighting (from moderately dark to bright) and weather conditions (sunny, cloudy, hazy), making detection particularly challenging. The current release contains 33,636 images in total, including 28,103 images with smoke and 31,975 smoke instances, annotated in YOLO format by Pyronear volunteers. Unlike `FASDD_UAV` and D-Fire, Pyro-SDIS primarily captures early-stage wildfire smoke, but the raw data include many near-duplicate frames and some annotation inconsistencies, so additional preprocessing is required before use in this study.

4.1.3. Dataset Construction

Two derived datasets are constructed for the experiments. The *Early Smoke* dataset merges `FASDD_UAV` and D-Fire, then applies near-duplicate removal and smoke-focused

https://doi.org/10.3390/fire9020092

filtering, while a separately deduplicated *Pyro-SDIS* split is built from the raw Pyro-SDIS release without additional smoke filtering.

Near-duplicate frames are removed to reduce metric inflation using a perceptual hashing (pHash) procedure based on low-frequency discrete cosine transform (DCT) features. Each image I is converted to grayscale and resized to 32×32 pixels using bilinear interpolation, producing a normalized representation I' that suppresses high-frequency variation. A 2D DCT is applied to I', and only the top-left 8×8 coefficients are retained because they capture the dominant low-frequency structure. Let $D = \{C(u,v) \mid 0 \leq u, v < 8\}$ denote this block. Then, the median value m of all coefficients in D is computed, and a 64-bit binary hash is generated by setting

$$h(u,v) = \begin{cases} 1, & C(u,v) > m, \\ 0, & C(u,v) \leq m, \end{cases}$$

followed by flattening the result in row-major order. Similarity between two images is evaluated via the Hamming distance between their hashes:

$$d(H_1, H_2) = \sum_{k=1}^{64} \mathbf{1}\Big(H_1^k \neq H_2^k\Big).$$

Images with $d(H_1, H_2) \leq 3$ are considered duplicates in the Early Smoke dataset, and $d(H_1, H_2) \leq 10$ flags duplicates within Pyro-SDIS. These empirically chosen thresholds, validated through extensive visual inspection, reliably remove near-identical frames—particularly those from consecutive UAV captures or repeated monitoring scenes—without discarding distinct examples.

After deduplication, additional filtering is applied to the Early Smoke dataset to isolate early-smoke instances. Pyro-SDIS is already curated for early-smoke detection and does not include the large-fire imagery commonly observed in FASDD–UAV and D-Fire; therefore, no additional smoke-specific filtering is applied to that split. This filtering procedure is intended to remove large-fire scenes while retaining images in which smoke is faint, spatially limited, and indicative of early ignition.

First, the annotated smoke region is required to occupy less than 2% of the total image area. Visual inspection of FASDD–UAV and D-Fire suggests that genuine early-smoke plumes typically occupy well under 1%; however, a 2% threshold is adopted to preserve coverage while excluding larger fire events. Second, each image is required to contain exactly one bounding box, consistent with the single ignition source expected in early-stage wildfires; images with multiple boxes are frequently associated with larger or fragmented fire scenes. Only images annotated with the `smoke` bounding-box label are retained, as visible flame is uncommon during the earliest ignition stages. The dataset is further restricted to high-resolution images with a longer side exceeding 1080 px, ensuring that multi-resolution composite generation can be performed without sacrificing spatial detail. These filtering steps yield a tightly curated Early Smoke set that reflects genuine early-smoke conditions and forms the foundation for evaluating fine-grained smoke detection.

4.1.4. Dataset Description and Statistics

The resulting datasets after the construction steps are summarized in Section 4.1. The filtered Early Smoke subset drawn from FASDD–UAV and D-Fire contains 936 images in which smoke is small (<2% of image area), faint, and set against varied backgrounds. The deduplicated Pyro-SDIS subset contains 3523 unique images captured by tower-mounted multi-camera systems in forested settings, where smoke typically appears distant and

https://doi.org/10.3390/fire9020092

low-contrast under diverse lighting and weather conditions; all instances are annotated in YOLO format by Pyronear volunteers. Accordingly, two datasets are constructed:

1. **Early Smoke Dataset**: 936 carefully filtered early-smoke images drawn from FASDD–UAV and D-Fire, emphasizing small, faint, low-contrast plumes with $<2\%$ image area under varying illumination and high background complexity.
2. **Pyro-SDIS Dataset**: 3523 deduplicated images from tower-mounted camera systems in predominantly forested scenes, containing far-distant, low-contrast smoke under diverse lighting and weather conditions, with manually curated YOLO annotations by Pyronear volunteers.

The two datasets serve complementary roles in our evaluation. The Early Smoke Dataset provides clean, controlled early-smoke instances tailored for training and detailed analysis of model behavior on canonical early-stage plumes. The Pyro-SDIS subset introduces a meaningful distribution shift in terms of hardware, geography, appearance, and atmospheric conditions, making it well suited for testing robustness and generalization under realistic deployment scenarios. Qualitative examples for each dataset are illustrated in Figures 4 and 5, respectively.

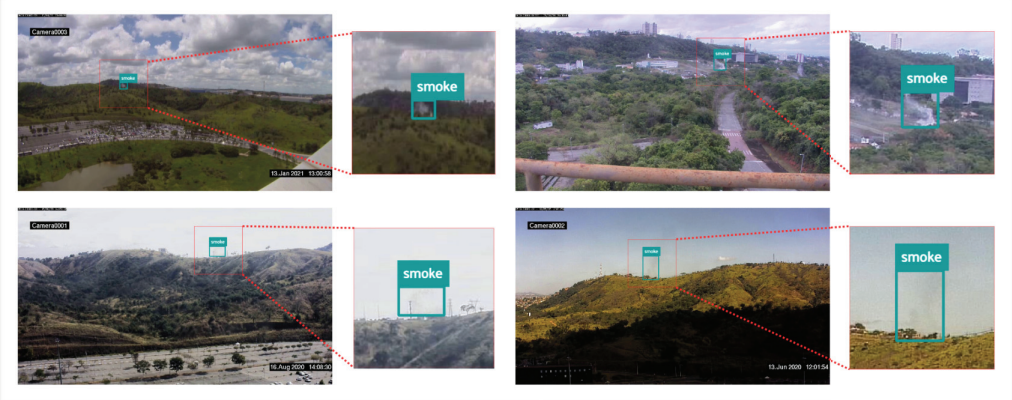


Figure 4. Qualitative examples from the Early Smoke dataset.

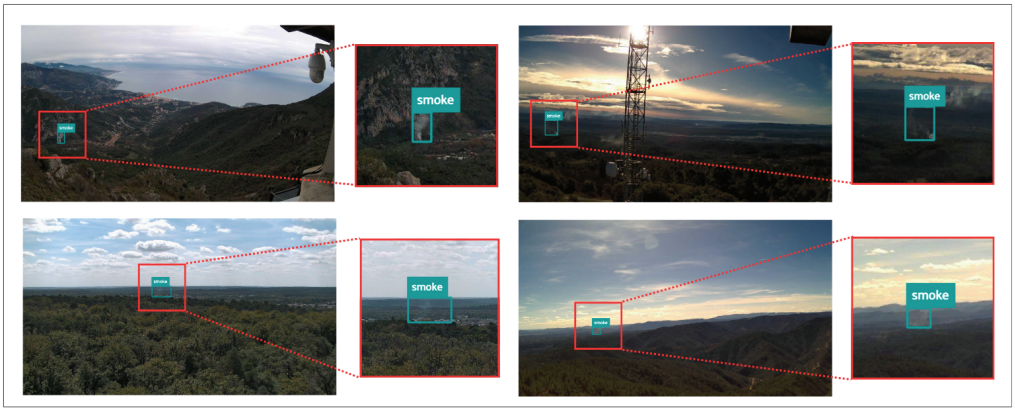


Figure 5. Qualitative examples from the Pyro-SDIS dataset.

4.2. Training for Early Smoke Detection

All experiments are conducted on a single workstation environment; Table 1 summarizes the hardware and software configuration used consistently across detector and classifier training to ensure reproducibility.

https://doi.org/10.3390/fire9020092

Table 1. Hardware and software environment for all experiments.

Component	Specification	Manufacturer (City, Country)
GPU	NVIDIA RTX A5000 (24 GB)	NVIDIA Corporation (Santa Clara, CA, USA)
CPU & Memory	Intel Core i5-14400F; 64 GB RAM	Intel Corporation (Santa Clara, CA, USA)
OS	Ubuntu 24.04.3 LTS	Canonical Ltd. (London, UK)
CUDA version	12.6	NVIDIA Corporation (Santa Clara, CA, USA)
Python version	Python 3.10.18	Python Software Foundation (Beaverton, OR, USA)
PyTorch version	2.7.1 + cu126	Meta Platforms, Inc. (Menlo Park, CA, USA)

The training pipeline proceeds in two stages. First, the composite-based detector is trained on both datasets using identical train/validation/test splits and the same composite-generation procedure used at inference to align distributions. Second, a post-hoc binary classifier is trained on detector proposals cropped from the original-resolution images to reduce false positives while preserving recall on faint, early-stage smoke.

4.2.1. Detector Training

The approach is evaluated using two curated datasets that focus on small, early-smoke instances. The first is the Early Smoke Dataset, consisting of 936 images obtained by filtering FASDD–UAV and D-Fire as described in Section 4.1. The second is the deduplicated Pyro-SDIS dataset, containing 3523 unique images.

All datasets are split at the image level into training, validation, and test sets using a 70/10/20 ratio. Test images are held out strictly for final evaluation and are never used for hyperparameter tuning; all model selection is performed using validation performance. Ground-truth annotations follow the standard YOLO format, consisting of class ID, box center, and box dimensions, all normalized to $[0, 1]$. During training and validation, each image is converted into a $W_{in} \times H_{in}$ composite using the procedure described in Section 3. The composite dimension is set to $W_{in} = H_{in} = 640$, selected empirically and chosen to match the widely used 640×640 detector input size in related object-detection pipelines.

Training is performed directly on composite images because inference in the deployed system is also conducted on composites. This alignment reduces distribution mismatch between training and deployment and avoids additional covariate shift introduced by composite generation. Test images, on the other hand, remain at their original resolution and are converted to composite form only at inference time. When creating a composite image for training and validation, the high-resolution crop is placed based on the ground-truth bounding box of the smoke instance, ensuring that the smoke region is fully contained within the crop window. This ensures that each annotated plume appears in both the global band and the high-resolution band of the composite, ensuring that the model learns from as many instances as possible. Under a fixed skyline-based inference policy, this strategy yields superior performance compared to training with skyline-based crop placement, which may not always capture the smoke region effectively during training. Table 2 presents an ablation study comparing these two training crop placement strategies under the same skyline-based inference setting.

Table 2. Ablation comparing training crop placement strategies under a fixed skyline-based inference policy.

Model	AP @0.5:0.95	AP @0.5	AP_{small}	F1	Precision	Recall
GT-based trained	0.420	0.852	0.212	0.832	0.881	0.787
Skyline-trained	0.397	0.845	0.113	0.812	0.832	0.793

To improve robustness to positional variation, three composite variants are generated for every training and validation image by shifting the crop horizontally so that the smoke

https://doi.org/10.3390/fire9020092

region appears near the left (≈25% of width), center (≈50%), or right (≈75%). This augmentation reflects deployment conditions, where the sky-adaptive logic may place the crop at different horizontal locations depending on the skyline estimate. In addition to positional augmentation, the default Ultralytics YOLO augmentations are applied to the training split only.

YOLOv8 was selected as the baseline detector because its performance is effectively on par with YOLOv9 in this evaluation (only marginal differences in AP and best-F1), while offering greater operational maturity for real-world deployment. In safety-critical wildfire detection, the cost of unexpected behavior in the field such as instability across environments, brittle exports, inconsistent runtime characteristics, or unanticipated integration issues can outweigh small metric gains. Compared to newer variants, YOLOv8 benefits from broader ecosystem adoption and longer exposure in production-like pipelines, which typically translates to better-tested tooling, more deployment references, and lower integration risk [51]. Given the negligible accuracy gap observed here, prioritizing deployment maturity and robustness is the more conservative choice for reliable end-to-end operation. Table 3 presents an ablation study comparing performances across YOLO variants under a consistent evaluation protocol.

Table 3. Detector comparison across YOLO variants using a consistent evaluation protocol.

Model	AP @0.50:0.95	AP @0.50	AP_{small}	F1	Precision	Recall
YOLOv12	0.400	0.843	0.127	0.838	0.902	0.782
YOLOv11	0.389	0.795	0.222	0.821	0.839	0.803
YOLOv10	0.379	0.814	0.153	0.806	0.775	0.840
YOLOv9	0.421	0.858	0.219	0.835	0.864	0.809
YOLOv8	0.420	0.852	0.212	0.832	0.881	0.787
YOLOv5su	0.409	0.825	0.102	0.811	0.802	0.819

For all experiments, the detector is trained for 100 epochs with a batch size of 16 and an initial learning rate of 10^{-2}, initialized from pretrained weights. The Early Smoke Dataset is optimized using AdamW, whereas the Pyro–SDIS dataset is optimized using stochastic gradient descent (SGD), reflecting the configurations that yielded the most stable convergence in each setting. These hyperparameters follow standard YOLO training practice and support reliable optimization across both small, low–contrast early–smoke instances and far–distance smoke scenes present in Pyro-SDIS.

During inference, the YCrCb thresholds for skyline estimation (Section 4.1.4) are fixed to empirically tuned values that stabilize skyline detection on the training images. Inputs are assumed to be RGB images. The sky mask retains pixels whose Cb and Cr values fall within luminance-adaptive ranges, where the lower and upper thresholds are obtained by linear interpolation across four luminance breakpoints, $Y = \{64, 128, 160, 192\}$. These breakpoint ranges were determined empirically by inspecting the Cb/Cr distributions of pixels corresponding to sky regions in the images used in our datasets. Specifically, the breakpoint values are set to Cb $\in [115, 162]$ and Cr $\in [110, 132]$ at $Y = 64$, Cb $\in [129, 186]$ and Cr $\in [96, 124]$ at $Y = 128$, Cb $\in [121, 183]$ and Cr $\in [93, 130]$ at $Y = 160$, and Cb $\in [118, 168]$ and Cr $\in [96, 136]$ at $Y = 192$. For a pixel with luminance Y between adjacent breakpoints, each of $\tau_{\text{Cb}}^{\min}(Y)$, $\tau_{\text{Cb}}^{\max}(Y)$, $\tau_{\text{Cr}}^{\min}(Y)$, and $\tau_{\text{Cr}}^{\max}(Y)$ is computed via piecewise linear interpolation between the corresponding endpoint values, and the pixel is labeled as sky if $\tau_{\text{Cb}}^{\min}(Y) \leq \text{Cb} \leq \tau_{\text{Cb}}^{\max}(Y)$ and $\tau_{\text{Cr}}^{\min}(Y) \leq \text{Cr} \leq \tau_{\text{Cr}}^{\max}(Y)$. A visualization of these thresholds is shown in Figure 6. The acceptance-ratio threshold is set to $\text{Th}_{sky} = 5.0$, and the skyline-estimation downsampling factor is set to $\alpha = 0.25$. These settings remain constant across all reported experiments.

https://doi.org/10.3390/fire9020092

When applied to the early smoke dataset, the skyline estimation procedure accurately identifies the presence of sky in 75% of the images. In cases where the skyline is correctly detected, the resulting crop placement effectively captures the smoke region within the high-resolution band of the composite, leading to improved detection performance. When the skyline is not detected, the fallback strategy of centering the crop vertically still allows for reasonable coverage of the smoke region, albeit with slightly reduced performance compared to successful skyline detection.

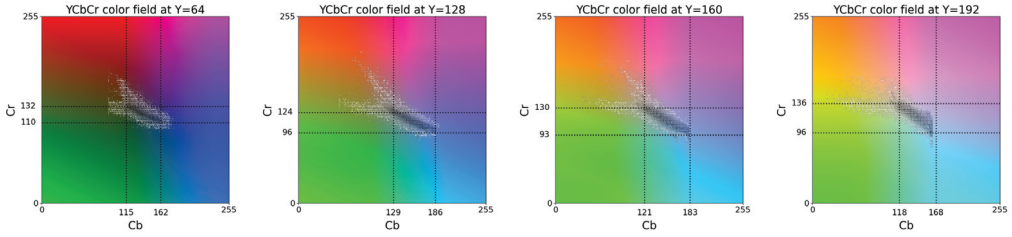


Figure 6. YCbCr planes at the four luminance breakpoints used to define the luminance-adaptive chrominance thresholds.

4.2.2. Training for Early Smoke Classification

To refine detector outputs and suppress low-confidence false positives, a binary classifier is trained on localized image crops derived from detector predictions. The classifier operates on 224×224 patches extracted from the original images and predicts whether each patch contains wildfire smoke.

The classifier is trained on the training splits of both the Early Smoke and Pyro-SDIS datasets. For each image, the same composite-generation pipeline used during detector inference is applied to maintain consistency between detector and classifier inputs. After the composite-based detector is trained, it is run on the training portions of both datasets and the resulting predicted bounding boxes are collected. These detector-generated proposals are then converted into 224×224 crops extracted from the original-resolution images and centered on each predicted region. By using detector proposals rather than ground-truth boxes, the classifier is trained on the same type of inputs encountered during deployment. This process yields a diverse set of localized patches containing both true smoke plumes and detector-induced false positives.

Each crop is assigned a label based on its IoU with the ground-truth smoke annotations. Crops with IoU > 0.5 are labeled as foreground (true smoke), whereas crops with IoU $< 10^{-6}$ are labeled as background (false positives). Crops with intermediate IoU values are discarded to mitigate ambiguity in supervision. After labeling, all crops are randomly partitioned into a 70/10/20 train/validation/test split, with splitting performed after crop generation to prevent leakage across splits. All classifier crops are extracted directly from the original-resolution images rather than from the composite representation. Cropping windows are centered on the detector's predicted bounding boxes, with appropriate boundary handling applied when detections occur near image borders. The resulting crop dataset is class-imbalanced, as detector-generated proposals contain substantially more background patches than true smoke instances. To mitigate this imbalance, the foreground samples are augmented until the numbers of positive and negative crops are approximately matched. The augmentation pipeline applies horizontal flips with probability 0.5 and introduces mild geometric perturbations, including small translations, slight scale variations, and rotations up to 15°. These transformations increase sample diversity while preserving the visual characteristics of early-stage smoke. Augmentation is applied only to the training split to keep the validation and test distributions unchanged.

https://doi.org/10.3390/fire9020092

For all classifier experiments, a YOLOv8-s classification model is trained, initialized from the pretrained yolov8s-cls.pt weights for 100 epochs with a batch size of 32 and an initial learning rate of 10^{-2}. The model is optimized using the AdamW optimizer, which provided stable convergence on the balanced foreground–background crop dataset derived from detector proposals. These hyperparameters follow standard YOLO classification practice and are sufficient to train a robust classifier capable of discriminating true smoke plumes from detector false positives.

4.3. Evaluation Metrics

Detection performance is evaluated using standard object detection metrics. Precision measures the fraction of predicted detections that are correct, while recall measures the fraction of ground-truth objects that are successfully detected. Both are defined as below where TP, FP, and FN denote true positives, false positives, and false negatives, respectively. The F1 score, which summarizes precision and recall into a single harmonic-mean metric.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \qquad \text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Average Precision (AP) quantifies detection performance by integrating the precision–recall curve at a specified IoU threshold. Experiments report AP at fixed thresholds (e.g., AP @0.5) as well as the COCO-style AP averaged over IoU thresholds from 0.50 to 0.95 in increments of 0.05. mAP is computed as the mean of AP values across the specified IoU thresholds and is used as the primary detection metric.

In addition, size-specific COCO metrics are also reported, including AP_{small}, $\text{AP}_{\text{medium}}$, and AP_{large}. These categories correspond to objects with areas smaller than 32^2, between 32^2 and 96^2, and larger than 96^2 pixels, respectively. These metrics provide insight into detection performance across different object scales.

5. Results and Discussion

For completeness, performance is evaluated on the Early Smoke and Pyro–SDIS datasets using (i) standard scientific metrics and (ii) a deployment setting with a fixed confidence threshold and uncertainty-band refinement. The baseline is a YOLOv8s detector trained on 640×640 letterboxed inputs without composite processing.

5.1. Overall Detection Performance

As shown in Table 4, our pipeline improves performance across all AP metrics, with the most notable gains in AP @0.5 and AP_{small} on the Early Smoke Dataset. The proposed method increases AP @0.5 from 0.818 to 0.852, resulting in a net improvement of 3.4%. The enhancement is most pronounced for small objects: AP_{small} increases from 0.054 to 0.212, representing nearly a fourfold improvement in the smallest-object regime. These improvements are more clearly reflected in the curves. As shown in Figure 7, the recall curves indicate relative gains of up to 60% in the high-confidence region (0.7–0.9). This behavior yields an overall increase of approximately 20% in average recall. These recall gains are accompanied by an expected reduction in precision of approximately 10% over the same confidence range. This trade-off is considered acceptable given the substantially larger recall improvement and the central importance of recall in early-smoke detection tasks. Additional evidence is provided by the ΔF1 curve, which indicates a net F1 increase of approximately 2%.

https://doi.org/10.3390/fire9020092

Table 4. Detection performance on the Early Smoke dataset.

Model	AP @0.5:0.95	AP @0.5	AP_{small}	F1	Precision	Recall
Baseline	0.374	0.818	0.054	0.789	0.852	0.734
Ours	0.420	0.852	0.212	0.832	0.881	0.787

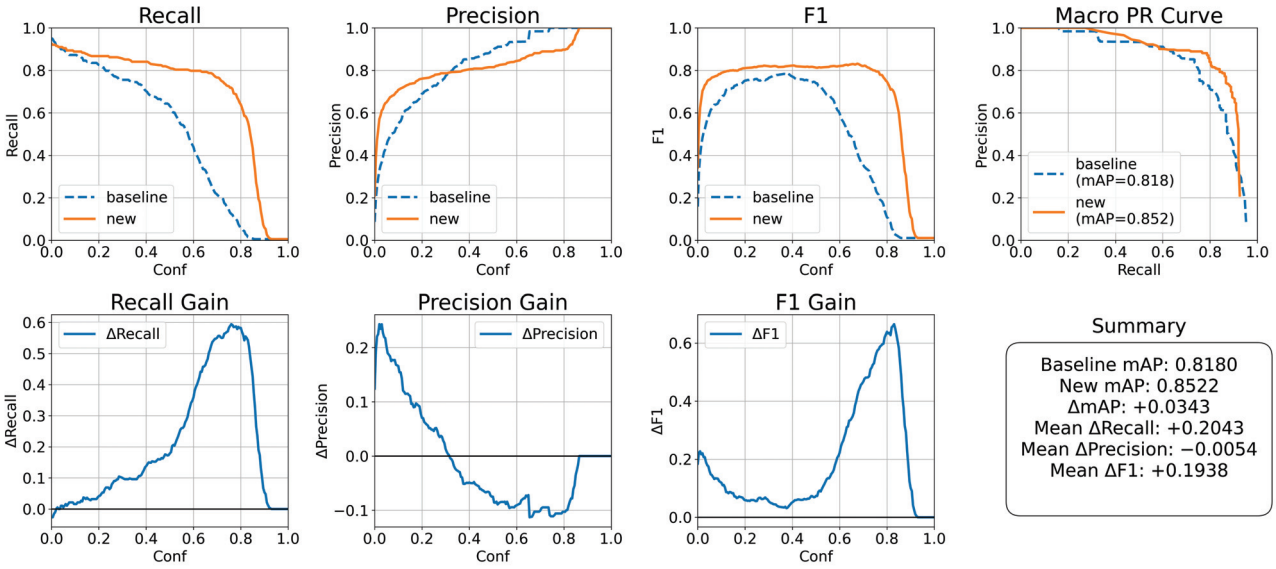


Figure 7. Comparison of baseline vs. composite detector performance on the Early Smoke dataset.

Beyond the baseline comparison, evaluations against other mainstream detector architectures indicate that the proposed pipeline consistently outperforms competing alternatives. As summarized in Table 5, the proposed method achieves the highest AP @0.5:0.95, AP @0.5, and AP @0.75. It also attains the best performance across all size-specific AP metrics. YOLOv5su exhibits strong AP @0.5 performance (0.839) but yields a comparatively lower AP_{small}. By contrast, YOLOv10s attains AP_{small} performance comparable to the proposed method (0.210) but underperforms on AP @0.5. The proposed method delivers balanced and superior results across metrics, supporting the effectiveness of composite-based training for early-stage wildfire smoke detection.

Table 5. Comparison against other mainstream detectors on the Early Smoke dataset.

Model	AP @0.5:0.95	AP @0.5	AP_{small}	F1	Precision	Recall
YOLOv5su	0.384	0.839	0.136	0.823	0.844	0.803
YOLOv8s	0.374	0.818	0.054	0.789	0.852	0.734
YOLOv9s	0.321	0.776	0.077	0.776	0.844	0.718
YOLOv10s (SGD)	0.367	0.804	0.210	0.785	0.816	0.755
YOLOv11s	0.379	0.792	0.134	0.770	0.852	0.702
Ours	0.420	0.852	0.212	0.832	0.881	0.787

To complement the quantitative results, Figures 8 and 9 present qualitative examples from the Early Smoke dataset. These examples illustrate the characteristic difficulty of the task. Smoke plumes are extremely small, visually faint, and frequently embedded in complex backgrounds. In many cases, baseline detectors either miss the plume entirely or produce low-confidence predictions. In contrast, the composite-trained model yields more consistent detections with higher confidence. This effect is most evident in scenes where the smoke occupies only a few dozen pixels. These examples provide visual evidence for the improvements reported in Table 4.

https://doi.org/10.3390/fire9020092

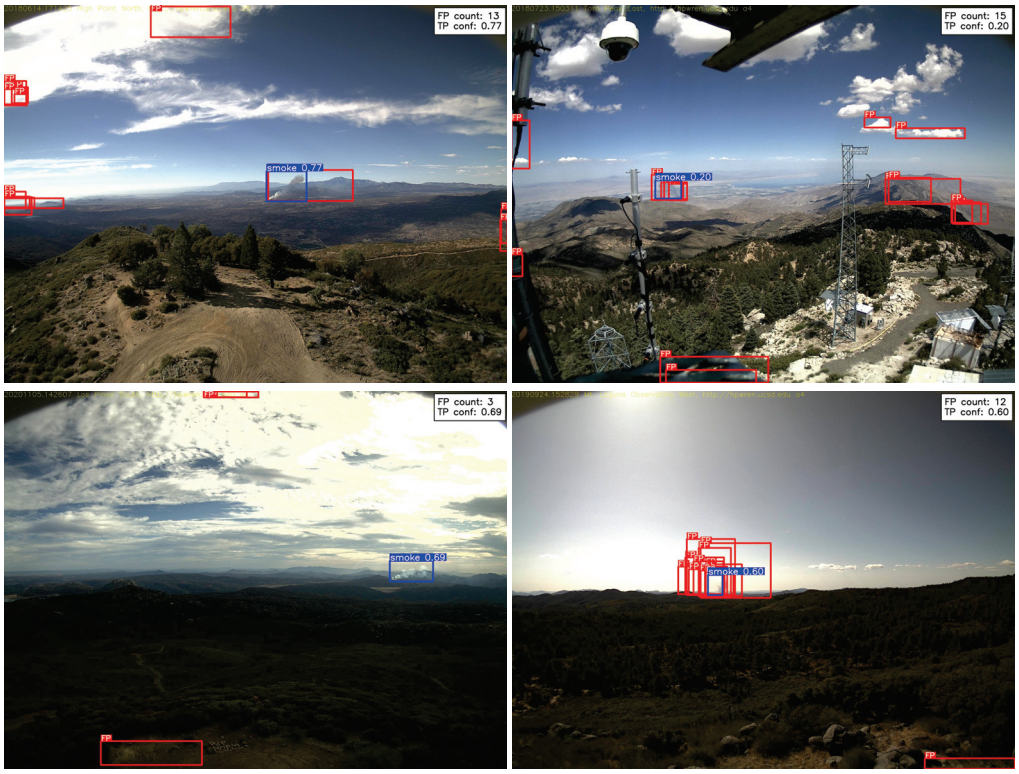


Figure 8. Qualitative detection examples on the Early Smoke dataset using baseline YOLOv8s detection. Bounding boxes are color-coded by detection outcome: blue denotes true positives and red denotes false positives.

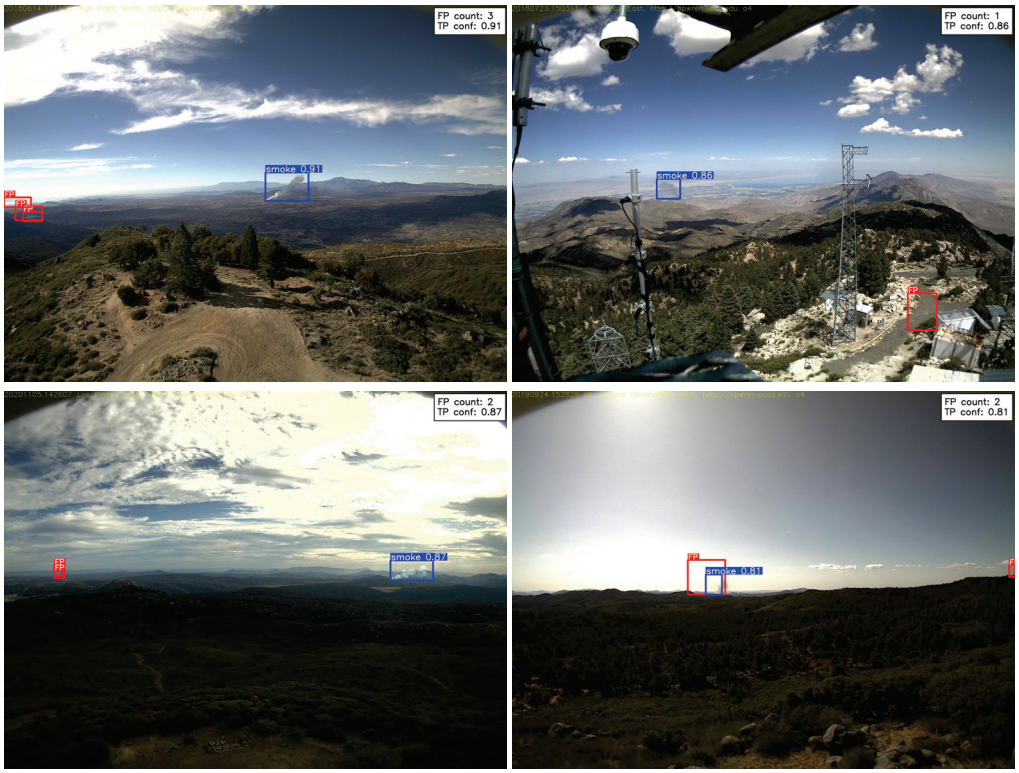


Figure 9. Qualitative detection examples on the Early Smoke dataset using the composite-trained model. Composite training improves confidence stability and plume localization for small smoke plumes.

https://doi.org/10.3390/fire9020092

Collectively, these results demonstrate that the proposed pipeline effectively enhances detection performance on small smoke objects that are critical for early wildfire detection. By leveraging multi-resolution compositing during training, the detector learns to better capture fine-grained smoke features while retaining global context, leading to substantial improvements in both recall and AP metrics, particularly for the smallest smoke plumes that are most challenging to detect. This indicates that the pipeline is especially effective under the difficult visual conditions of early-smoke detection, where standard single-resolution models struggle to preserve sufficient fine-scale structure.

5.2. Deployment-Only Evaluation (Two-Stage Classifier)

In deployment settings, the detector operates under a fixed confidence threshold that balances false alarms and missed detections. Early-smoke predictions often lie near the decision boundary, where confidence values fluctuate and standard thresholding becomes unstable. To address this issue, a two-stage classifier is introduced to refine only detections whose confidence falls within a predefined uncertainty band. This refinement is applied exclusively at deployment time and does not affect the detector's scientific evaluation reported in the previous section.

Confidence Threshold Selection

Figure 10 shows the confidence–recall and confidence–precision curves computed for the Early Smoke dataset. The detector exhibits a sharp drop in recall only below a confidence of approximately 0.2, making $\mathrm{Th}_{conf} = 0.2$ a suitable operating threshold for recall-oriented deployment. The region to which the classifier is applied must be minimized, since the refinement stage introduces substantial computational overhead. Moreover, precision is relatively unstable in the low-confidence range because smoke plumes are often extremely small and visually faint. Accordingly, the uncertainty band for classifier refinement is selected by analyzing the confidence-score distribution of true positive detections around confidence values near 0.2. It is empirically observed that a large proportion of true detections fall within the 0.1–0.3 interval, reinforcing this range as the appropriate target for refinement. Hence, we set $\mathrm{Th}_{low} = 0.1$ and $\mathrm{Th}_{high} = 0.3$ as the uncertainty band thresholds. These values balance recall preservation with minimal disruption to high-confidence detector outputs, while still enabling the classifier to correct uncertain cases.

Table 6 reports deployment-setting performance on the Early Smoke dataset under a fixed confidence threshold of $\mathrm{Th}_{conf} = 0.2$. The multiresolution detector substantially outperforms the baseline YOLOv8 across all metrics, including $\mathrm{AP}_{50:95}$, AP_{75}, and AP_{S}, establishing a stronger operating point under fixed-threshold filtering. When the two-stage classifier is applied to detections within the 0.1–0.3 uncertainty band, additional gains are observed. These gains are driven by more reliable handling of borderline predictions, which are most sensitive to hard-threshold decisions.

Notably, this selective refinement recovers true positives that would otherwise be discarded when using a single detector score for thresholding. In operational settings, detections near Th_{conf} often include visually ambiguous but correct early-smoke instances; the classifier provides an additional verification signal that can promote these candidates above the decision threshold while suppressing low-confidence false positives. As a result, the two-stage system improves overall AP and especially strengthens small-object performance without requiring a lower global threshold, making it more closely aligned with deployment constraints where fixed-threshold decisions are necessary. Figures 11 and 12 present qualitative examples comparing baseline detections against the full pipeline with multiresolution detection and two-stage classification.

https://doi.org/10.3390/fire9020092

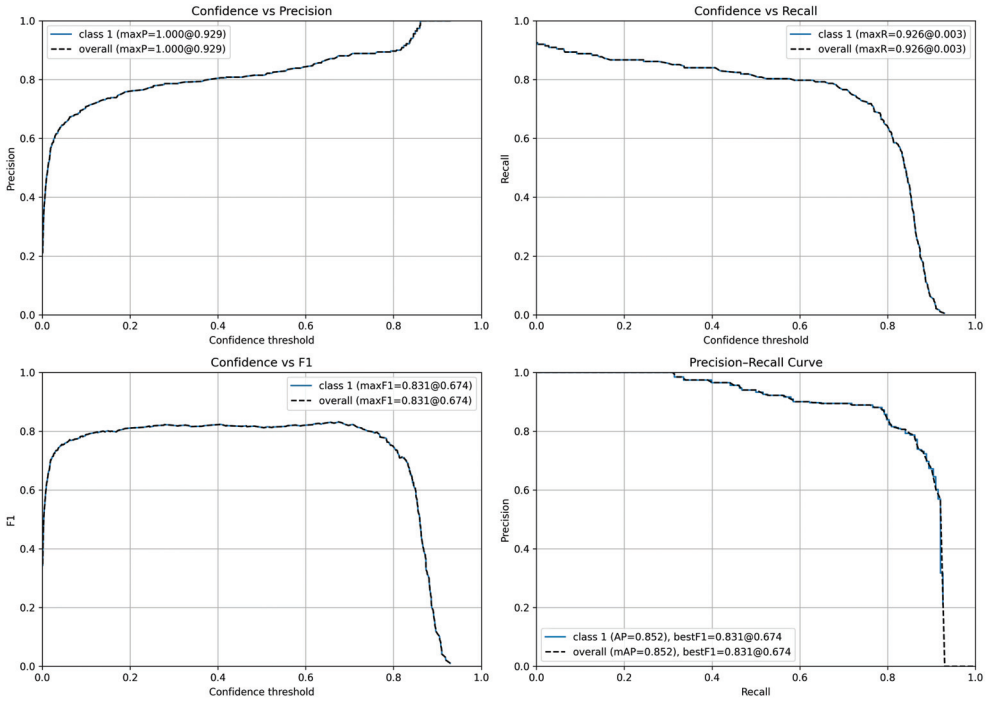


Figure 10. Confidence–recall and confidence–precision curves for the Early Smoke dataset. The selected deployment threshold of $\mathrm{Th}_{conf} = 0.2$ preserves recall while the 0.1–0.3 band defines the refinement region for the two-stage classifier.

Table 6. Deployment evaluation on the Early Smoke dataset at a fixed threshold of $\mathrm{Th}_{conf} = 0.2$, compared with the YOLOv8s baseline detector. The two-stage classifier refines detections in the 0.1–0.3 confidence interval.

Experiment	AP @ 0.50:0.95	AP @ 0.50	AP_{small}	F1	Precision	Recall
Baseline	0.345	0.760	0.037	0.789	0.852	0.734
Multiresolution	0.404	0.813	0.188	0.832	0.881	0.787
Multiresolution + Classifier	0.410	0.820	0.188	0.832	0.881	0.787

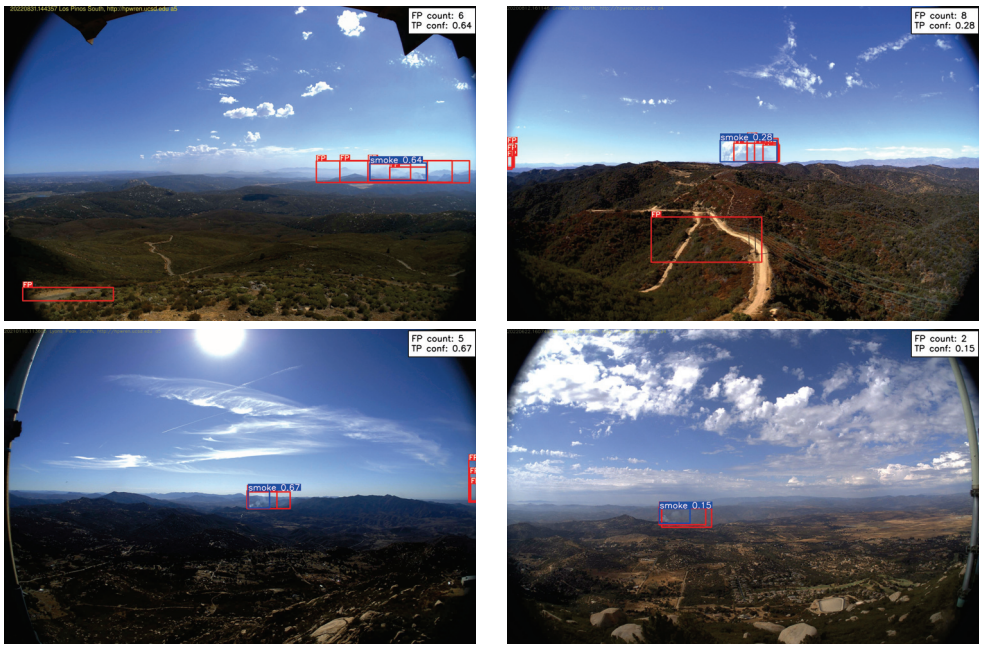


Figure 11. Qualitative ablation examples using the baseline detector. Bounding boxes are color-coded by outcome: blue denotes true positives and red denotes false positives.

https://doi.org/10.3390/fire9020092

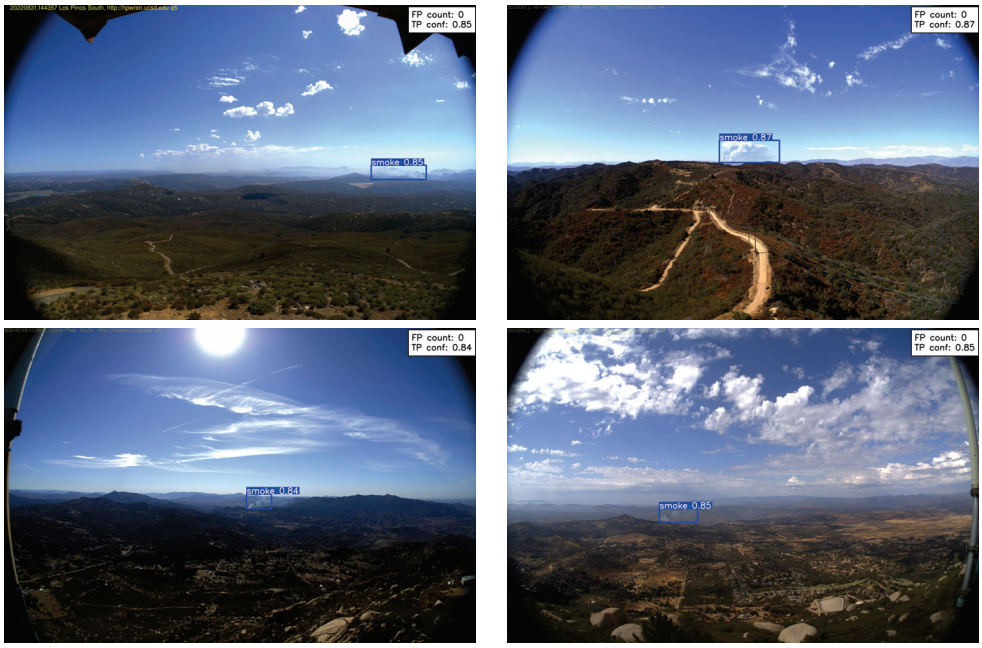


Figure 12. Qualitative ablation examples using the full pipeline.

5.3. Ablation Study

To quantify the contribution of each module and to characterize the accuracy–efficiency trade-off of the complete system, we conduct an ablation study in which the multiresolution detector and the two-stage classifier are enabled individually and in combination. Such analysis is required to attribute performance gains to specific design choices and to identify the primary sources of additional latency under full deployment. In particular, because the second-stage classifier introduces extra computation, the ablation results provide empirical justification for a selective, event-driven invocation strategy, in which the classifier is applied only to ambiguous detector outputs rather than uniformly across all frames, thereby limiting overhead during background-dominated surveillance. Table 7 reports an ablation study analyzing the individual and combined effects of the multiresolution detector and the two-stage classifier on the Early Smoke dataset. The baseline YOLOv8 model exhibits very low performance on small objects ($AP_S = 0.037$), reflecting the extreme difficulty of detecting early-stage smoke plumes that occupy only a few pixels and lack strong visibility.

Introducing the two-stage classifier alone yields only marginal changes (slight gains in AP_{50} and AP_M) and does not lift AP_S, confirming that confidence refinement cannot recover missed tiny plumes without better spatial evidence. In contrast, enabling the multiresolution representation produces the dominant improvement: AP_S increases from 0.037 to 0.188 and higher-IoU scores (AP_{75}) also rise, indicating sharper localization as well as recall gains. Combining multiresolution detection with the two-stage classifier yields the best overall performance ($AP_{50:95} = 0.410$), with the classifier acting as a lightweight stabilizer in the 0.1–0.3 band rather than as a primary source of recall. These results underline that the high-resolution composite is the key driver for faint-smoke sensitivity, while the selective second stage mainly curbs borderline false alarms without suppressing true positives.

In addition to accuracy, Table 8 also reports end-to-end inference time for each configuration. While the full pipeline introduces higher worst-case latency due to the additional multiresolution processing and the conditional second-stage classifier, this cost is not incurred uniformly across frames in practical surveillance settings. The classifier is applied only to detections within a narrow confidence band (e.g., 0.1–0.3), which occur infrequently given the rarity of early smoke events and the predominance of background-only frames. For example, over a stream of 10,000 frames, if the classifier is triggered on only 10% of

https://doi.org/10.3390/fire9020092

frames, the amortized per-frame runtime is approximately 9.5 ms, compared to 16.3 ms in the worst case. As a result, the most expensive component of the pipeline is triggered sparsely and in an event-driven manner, yielding an expected (amortized) runtime that remains close to that of the single-stage detector over long video streams. This design allows the system to prioritize sensitivity and robustness when ambiguous signals arise, while preserving efficient operation during normal background monitoring.

Table 7. Ablation study on the Early Smoke dataset evaluating the effects of multiresolution representation and the two-stage classifier. The classifier refines detections in the 0.1–0.3 confidence interval.

Multiresolution	Classifier	AP @ 0.50:0.95	AP @ 0.50	AP_{small}	F1	Precision	Recall
X	X	0.345	0.760	0.037	0.789	0.852	0.734
X	O	0.343	0.768	0.037	0.806	0.804	0.809
O	X	0.404	0.813	0.188	0.832	0.881	0.787
O	O	**0.410**	**0.820**	**0.188**	0.832	0.881	0.787

Table 8. Inference-time performance for the ablation settings in Table 7.

Multiresolution	Classifier	Inference Time (ms)	FPS
X	X	3.168	315.66
X	O	7.515	133.07
O	X	5.825	171.67
O	O	11.708	85.41

5.4. Generalizability Verification

To assess generalizability, the proposed pipeline is evaluated on Pyro-SDIS under an identical operating threshold and refinement band. Table 9 shows that the proposed model achieves the highest AP @0.5:0.95, AP @0.5, and AP @0.75 among mainstream detectors, indicating that the multiresolution composite and skyline-guided placement transfer beyond the primary dataset. Scale-specific gains on small and medium objects suggest that the same mechanisms that enhance faint-plume sensitivity on Early Smoke also improve performance on Pyro-SDIS, while AP for size-specific objects remains competitive. This pattern indicates robustness to distribution shifts in background, scene layout, and smoke appearance without dataset-specific tuning.

Table 9. Comparative performance of multiresolution pipeline on the Pyro-SDIS dataset.

Model	AP @0.5:0.95	AP @0.5	AP @0.75	AP_{small}	F1	Precision	Recall
YOLOv5su	0.419	0.714	0.445	0.400	0.700	0.683	0.719
YOLOv8s	0.418	0.737	0.423	0.405	0.729	0.745	0.714
YOLOv9s	0.406	0.703	0.405	0.361	0.687	0.698	0.676
YOLOv10s	0.413	0.713	0.432	0.391	0.699	0.750	0.655
YOLOv11s	0.410	0.718	0.417	0.367	0.691	0.691	0.690
Ours	**0.436**	**0.749**	**0.447**	**0.410**	**0.726**	**0.777**	**0.681**

Table 10 reports deployment-setting performance on Pyro–SDIS using the same operating threshold ($Th_{conf} = 0.2$) and the same 0.1–0.3 refinement band. The multiresolution detector improves $AP_{50:95}$, AP_{50}, and AP_S over the baseline, and the selective two-stage pass adds incremental gains, reaching $AP_{50:95} = 0.415$, $AP_{50} = 0.700$, and $AP_S = 0.390$. The progression mirrors the primary dataset: multiresolution delivers most of the lift, and the classifier offers a modest stability boost, indicating that the deployment recipe transfers without dataset-specific tuning. In addition to the performance metrics, Table 11 reports the corresponding runtime performance.

https://doi.org/10.3390/fire9020092

Table 10. Deployment-setting performance on the Pyro-SDIS dataset under a fixed decision threshold of $\mathrm{Th}_{conf} = 0.2$. The classifier is applied only to detections within the 0.1–0.3 confidence band.

Multiresolution	Classifier	AP @0.50:0.95	AP @0.50	AP @0.75	AP_{small}	F1	Precision	Recall
X	X	0.386	0.664	0.404	0.365	0.724	0.758	0.692
X	O	0.401	0.667	0.424	0.375	0.734	0.708	0.762
O	X	0.410	0.691	0.426	0.386	0.726	0.777	0.681
O	O	0.416	0.703	0.431	0.394	0.726	0.777	0.681

Table 11. Runtime performance for the deployment settings in Table 10.

Multiresolution	Classifier	Average Inference Time (ms)	FPS
X	X	3.238	308.83
X	O	6.741	148.35
O	X	5.630	177.62
O	O	13.818	72.37

Failure Analysis

While the proposed pipeline improves detection performance, several failure modes persist. The most common failures occur when smoke plumes are extremely faint, very small, or low-contrast, causing them to blend into complex backgrounds and leading to missed detections even with multiresolution input. Errors also arise when skyline estimation becomes unreliable under challenging conditions—such as cloud occlusion, strong lighting variation, or foreground occlusions—resulting in suboptimal crop placement that excludes critical plume details. In addition, the two-stage classifier can occasionally reject true positives when the smoke appearance is highly ambiguous, and false positives may occur when clouds or other scene elements exhibit smoke-like color and texture, particularly under overcast skies. Finally, ambiguous plume boundaries can produce inconsistent annotations, which complicates evaluation and can amplify apparent errors even when predictions are visually plausible. Figure 13 illustrates several representative failure cases, highlighting the ongoing challenges in accurately detecting early-stage wildfire smoke under diverse and difficult visual conditions.

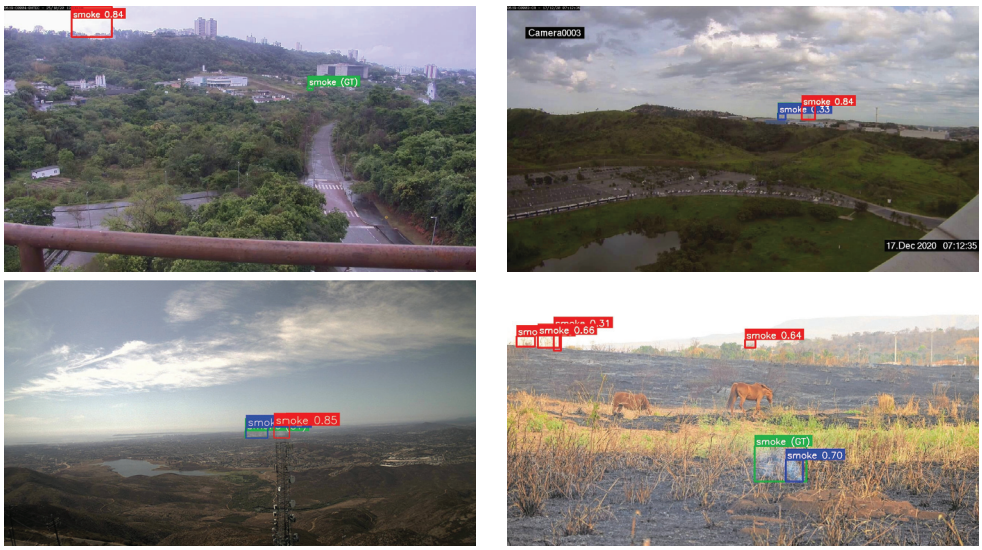


Figure 13. Qualitative failure cases for early-smoke detection. Green denotes ground truth, red denotes false positives, and blue denotes true positives.

https://doi.org/10.3390/fire9020092

6. Conclusions

The proposed multi-resolution pipeline addresses the central challenge of detecting faint, small-scale wildfire smoke. By embedding a high-resolution crop within a standard square input and aligning it with skyline-guided placement, the detector recovers fine-grained plume structure while retaining global context. Our approach outperforms the baseline on the Early Smoke dataset, improving AP @0.5:0.95 by +4.6 percentage points, mAP @0.5 by +3.4 percentage points, precision by +2.9 percentage points, recall by +5.3 percentage points, and F1 score by +4.3 percentage points. Ablation results show that this composite representation is the primary driver of the substantial AP gains on early-smoke instances, especially in AP_S, and that these benefits carry over to a secondary dataset without dataset-specific tuning. The selective two-stage classifier further stabilizes deployment behavior by refining detections only within a narrow uncertainty band, reducing false alarms near the operating threshold without altering scientific evaluation. The pipeline delivers measurable improvements in recall and localization for early smoke while keeping latency within practical bounds for field deployment. The combination of single-pass multiresolution detection, reversible coordinate remapping, and lightweight selective refinement provides a deployable balance between accuracy and efficiency for real-time wildfire monitoring.

However, several limitations remain. First, skyline estimation errors can lead to suboptimal placement of the high-resolution crop, potentially excluding critical plume regions. Because the skyline is derived using a simple edge-based heuristic, it is sensitive to cloud cover, illumination changes, and foreground occlusions. When the estimated skyline is inaccurate, the resulting crop can be misaligned, which may reduce detection performance. Second, although less frequent, smoke plumes may occur near the extreme lateral boundaries of the image and fall outside the high-resolution crop even when the skyline prior behaves as intended. While the global bottom-band view mitigates complete miss detections in such cases, the proposed pipeline may provide limited benefit because the plume is effectively observed only in the lower-resolution global context.

Future work will focus on improving the robustness and coverage of the multiresolution cropping strategy. A primary direction is to replace the edge-based skyline estimator with a more reliable, learning-based module such as lightweight segmentation module that is resilient to extreme conditions and can provide an uncertainty estimate to guide crop placement. In addition, we plan to mitigate boundary cases by adopting adaptive cropping policies—such as multiple high-resolution crops, dynamic crop widths, or lightweight tiling near image extremes while minimizing the computational cost such that salient plume regions are more consistently captured at higher resolution.

Author Contributions: Conceptualization, G.J. and B.M.; methodology, G.J. and B.M.; software, G.J.; validation, G.J.; formal analysis, G.J.; investigation, G.J.; resources, B.M.; data curation, G.J.; writing—original draft preparation, G.J.; writing—review and editing, G.J., T.-H.A. and B.M.; visualization, G.J.; supervision, B.M.; project administration, B.M. and T.-H.A.; funding acquisition, B.M. and T.-H.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Regional Innovation System & Education (RISE) through the Seoul RISE Center, funded by the Ministry of Education (MOE) and the Seoul Metropolitan Government (2025-RISE-01-019-04). In addition, Tae-Hyuk Ahn was supported by the National Science Foundation (NSF) under Grant No. 2430236.

Data Availability Statement: The curated Early Smoke dataset, the deduplicated Pyro-SDIS dataset, and the accompanying code will be released publicly at: https://github.com/vision-ai-lab/fire-detection (accessed on 9 January 2026).

https://doi.org/10.3390/fire9020092

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Arteaga, B.; Diaz, M.; Jojoa, M. Deep learning applied to forest fire detection. In *Proceedings of the 2020 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*; IEEE: New York, NY, USA, 2020; pp. 1–6.
2. Sun, J.; Qi, W.; Huang, Y.; Xu, C.; Yang, W. Facing the Wildfire Spread Risk Challenge: Where Are We Now and Where Are We Going? *Fire* **2023**, *6*, 228. [CrossRef]
3. Vasconcelos, R.N.; Franca Rocha, W.J.; Costa, D.P.; Duverger, S.G.; Santana, M.M.d.; Cambui, E.C.; Ferreira-Ferreira, J.; Oliveira, M.; Barbosa, L.d.S.; Cordeiro, C.L. Fire detection with deep learning: A comprehensive review. *Land* **2024**, *13*, 1696. [CrossRef]
4. Zhao, Y.; Ma, J.; Li, X.; Zhang, J. Saliency detection and deep learning-based wildfire identification in UAV imagery. *Sensors* **2018**, *18*, 712. [CrossRef]
5. Wu, H.; Li, H.; Shamsoshoara, A.; Razi, A.; Afghah, F. Transfer learning for wildfire identification in UAV imagery. In *Proceedings of the 2020 54th Annual Conference on Information Sciences and Systems (CISS)*; IEEE: New York, NY, USA, 2020; pp. 1–6.
6. Zhao, L.; Zhi, L.; Zhao, C.; Zheng, W. Fire-YOLO: A small target object detection method for fire inspection. *Sustainability* **2022**, *14*, 4930. [CrossRef]
7. Ahn, Y.; Choi, H.; Kim, B.S. Development of early fire detection model for buildings using computer vision-based CCTV. *J. Build. Eng.* **2023**, *65*, 105647. [CrossRef]
8. Huang, P.; Chen, M.; Chen, K.; Zhang, H.; Yu, L.; Liu, C. A combined real-time intelligent fire detection and forecasting approach through cameras based on computer vision method. *Process Saf. Environ. Prot.* **2022**, *164*, 629–638. [CrossRef]
9. Girshick, R. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*; IEEE: New York, NY, USA, 2015; pp. 1440–1448.
10. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767. [CrossRef]
11. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 213–229.
12. Lv, C.; Zhou, H.; Chen, Y.; Fan, D.; Di, F. A lightweight fire detection algorithm for small targets based on YOLOv5s. *Sci. Rep.* **2024**, *14*, 14104. [CrossRef]
13. Xiao, Z.; Wan, F.; Lei, G.; Xiong, Y.; Xu, L.; Ye, Z.; Liu, W.; Zhou, W.; Xu, C. Fl-yolov7: A lightweight small object detection algorithm in forest fire detection. *Forests* **2023**, *14*, 1812. [CrossRef]
14. Zhang, L.; Wang, M.; Ding, Y.; Bu, X. MS-FRCNN: A multi-scale faster RCNN model for small target forest fire detection. *Forests* **2023**, *14*, 616. [CrossRef]
15. Wang, Y.; Bashir, S.M.A.; Khan, M.; Ullah, Q.; Wang, R.; Song, Y.; Guo, Z.; Niu, Y. Remote sensing image super-resolution and object detection: Benchmark and state of the art. *Expert Syst. Appl.* **2022**, *197*, 116793. [CrossRef]
16. Boroujeni, S.P.H.; Razi, A.; Khoshdel, S.; Afghah, F.; Coen, J.L.; O'Neill, L.; Fule, P.; Watts, A.; Kokolakis, N.M.T.; Vamvoudakis, K.G. A comprehensive survey of research towards AI-enabled unmanned aerial systems in pre-, active-, and post-wildfire management. *Inf. Fusion* **2024**, *108*, 102369. [CrossRef]
17. Andrianarivony, H.S.; Akhloufi, M.A. Machine learning and deep learning for wildfire spread prediction: A review. *Fire* **2024**, *7*, 482. [CrossRef]
18. Peng, Y.; Wang, Y. Automatic wildfire monitoring system based on deep learning. *Eur. J. Remote Sens.* **2022**, *55*, 551–567. [CrossRef]
19. Afghah, F. Autonomous Unmanned Aerial Vehicle Systems in Wildfire Detection and Management-Challenges and Opportunities. In *Proceedings of the Dynamic Data Driven Application Systems*; Springer Nature: Cham, Switzerland, 2022.
20. Lelis, C.A.S.; Roncal, J.J.; Silveira, L.; De Aquino, R.D.G.; Marcondes, C.A.C.; Marques, J.; Loubach, D.S.; Verri, F.A.N.; Curtis, V.V.; De Souza, D.G. Drone-Based AI System for Wildfire Monitoring and Risk Prediction. *IEEE Access* **2024**, *12*, 139865–139882. [CrossRef]
21. Bailon-Ruiz, R.; Bit-Monnot, A.; Lacroix, S. Real-time wildfire monitoring with a fleet of UAVs. *Robot. Auton. Syst.* **2022**, *152*, 104071. [CrossRef]
22. Phillips, W., III; Shah, M.; da Vitoria Lobo, N. Flame recognition in video. *Pattern Recognit. Lett.* **2002**, *23*, 319–327. [CrossRef]
23. Gubbi, J.; Marusic, S.; Palaniswami, M. Smoke detection in video using wavelets and support vector machines. *Fire Saf. J.* **2009**, *44*, 1110–1115. [CrossRef]
24. Mambile, C.; Kaijage, S.; Leo, J. Application of Deep Learning in Forest Fire Prediction: A Systematic Review. *IEEE Access* **2024**, *12*, 190554–190581. [CrossRef]

https://doi.org/10.3390/fire9020092

25. Ghali, R.; Akhloufi, M.A. Deep learning approaches for wildland fires remote sensing: Classification, detection, and segmentation. *Remote Sens.* **2023**, *15*, 1821. [CrossRef]
26. Akagic, A.; Buza, E. Lw-fire: A lightweight wildfire image classification with a deep convolutional neural network. *Appl. Sci.* **2022**, *12*, 2646. [CrossRef]
27. Seydi, S.T.; Saeidi, V.; Kalantar, B.; Ueda, N.; Halin, A.A. Fire-Net: A Deep Learning Framework for Active Forest Fire Detection. *J. Sens.* **2022**, *2022*, 8044390. [CrossRef]
28. Ramos, L.; Casas, E.; Bendek, E.; Romero, C.; Rivas-Echeverría, F. Computer vision for wildfire detection: A critical brief review. *Multimed. Tools Appl.* **2024**, *83*, 83427–83470. [CrossRef]
29. Li, J.; Tang, H.; Li, X.; Dou, H.; Li, R. LEF-YOLO: A lightweight method for intelligent detection of four extreme wildfires based on the YOLO framework. *Int. J. Wildland Fire* **2023**, *33*, WF23044. [CrossRef]
30. Bhargav, R.; Singh, P. Efficient UAV-Based Forest Fire Detection Using CNN and YOLOv8 Integration. In *Proceedings of the 2025 6th International Conference on Recent Advances in Information Technology (RAIT)*; IEEE: New York, NY, USA, 2025; pp. 1–6.
31. Jonnalagadda, A.V.; Hashim, H.A. SegNet: A segmented deep learning based Convolutional Neural Network approach for drones wildfire detection. *Remote Sens. Appl. Soc. Environ.* **2024**, *34*, 101181. [CrossRef]
32. Khryashchev, V.V.; Larionov, R. Wildfire Segmentation on Satellite Images using Deep Learning. In *Proceedings of the 2020 Moscow Workshop on Electronic and Networking Technologies (MWENT)*; IEEE: New York, NY, USA, 2020; pp. 1–5.
33. Bouguettaya, A.; Zarzour, H.; Taberkit, A.M.; Kechida, A. A review on early wildfire detection from unmanned aerial vehicles using deep learning-based computer vision algorithms. *Signal Process.* **2022**, *190*, 108309. [CrossRef]
34. Ghali, R.; Akhloufi, M.A.; Mseddi, W.S. Deep learning and transformer approaches for UAV-based wildfire detection and segmentation. *Sensors* **2022**, *22*, 1977. [CrossRef]
35. Cheng, G.; Yuan, X.; Yao, X.; Yan, K.; Zeng, Q.; Xie, X.; Han, J. Towards large-scale small object detection: Survey and benchmarks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 13467–13488. [CrossRef]
36. Wang, J.; Yang, W.; Guo, H.; Zhang, R.; Xia, G.S. Tiny object detection in aerial images. In *Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR)*; IEEE: New York, NY, USA, 2021; pp. 3791–3798.
37. Jiang, L.; Yuan, B.; Du, J.; Chen, B.; Xie, H.; Tian, J.; Yuan, Z. MFFSODNet: Multiscale feature fusion small object detection network for UAV aerial images. *IEEE Trans. Instrum. Meas.* **2024**, *73*, 3381272. [CrossRef]
38. Hu, M.; Li, Z.; Yu, J.; Wan, X.; Tan, H.; Lin, Z. Efficient-lightweight YOLO: Improving small object detection in YOLO for aerial images. *Sensors* **2023**, *23*, 6423. [CrossRef]
39. Wang, G.; Chen, Y.; An, P.; Hong, H.; Hu, J.; Huang, T. UAV-YOLOv8: A small-object-detection model based on improved YOLOv8 for UAV aerial photography scenarios. *Sensors* **2023**, *23*, 7190. [CrossRef]
40. Ozge Unel, F.; Ozkalayci, B.O.; Cigla, C. The Power of Tiling for Small Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*; IEEE: New York, NY, USA, 2019.
41. Akyon, F.C.; Altinuc, S.O.; Temizel, A. Slicing aided hyper inference and fine-tuning for small object detection. In *Proceedings of the 2022 IEEE International Conference on Image Processing (ICIP)*; IEEE: New York, NY, USA, 2022; pp. 966–970.
42. Zhang, H.; Hao, C.; Song, W.; Jiang, B.; Li, B. Adaptive slicing-aided hyper inference for small object detection in high-resolution remote sensing images. *Remote Sens.* **2023**, *15*, 1249. [CrossRef]
43. Hao, C.; Zhang, H.; Song, W.; Liu, F.; Wu, E. Slinet: Slicing-aided learning for small object detection. *IEEE Signal Process. Lett.* **2024**, *31*, 790–794. [CrossRef]
44. Telçeken, M.; Akgun, D.; Kacar, S. An evaluation of image slicing and YOLO architectures for object detection in UAV images. *Appl. Sci.* **2024**, *14*, 11293. [CrossRef]
45. Muzammul, M.; Li, X.; Li, X. Enhancing Tiny Object Detection Using Guided Object Inference Slicing (GOIS): An efficient dynamic adaptive framework for fine-tuned and non-fine-tuned deep learning models. *Neurocomputing* **2025**, *640*, 130327. [CrossRef]
46. Chen, Z.; Chen, G. STTSBI: A Fast Inference Framework for Small Object Detection in Ultra-High-Resolution Images. In *Proceedings of the 2024 4th International Conference on Intelligent Technology and Embedded Systems (ICITES)*; IEEE: New York, NY, USA, 2024; pp. 129–135.
47. Koyun, O.C.; Keser, R.K.; Akkaya, I.B.; Töreyin, B.U. Focus-and-Detect: A small object detection framework for aerial images. *Signal Process. Image Commun.* **2022**, *104*, 116675. [CrossRef]
48. Wang, M.; Yue, P.; Jiang, L.; Yu, D.; Tuo, T.; Li, J. An open flame and smoke detection dataset for deep learning in remote sensing based fire detection. *Geo-Spat. Inf. Sci.* **2025**, *28*, 511–526. [CrossRef]
49. De Venâncio, P.V.A.; Rezende, T.M.; Lisboa, A.C.; Barbosa, A.V. Fire detection based on a two-dimensional convolutional neural network and temporal analysis. In *Proceedings of the 2021 IEEE Latin American Conference on Computational Intelligence (LA-CCI)*; IEEE: New York, NY, USA, 2021; pp. 1–6.

https://doi.org/10.3390/fire9020092

50. Pyronear Team. Pyro-SDIS Dataset. Hugging Face. 2024. Available online: https://huggingface.co/datasets/pyronear/pyro-sdis (accessed on 9 January 2026).
51. Sapkota, R.; Karkee, M. Ultralytics YOLO evolution: An overview of YOLO26, YOLO11, YOLOv8 and YOLOv5 object detectors for computer vision and pattern recognition. *arXiv* **2025**, arXiv:2510.09653.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

https://doi.org/10.3390/fire9020092

fire

MDPI

Article

Applying an Interpretable Deep Learning Model to Identify Wildfire-Prone Areas in Southwest China

Chenyu Ma [1], Siquan Yang [1,2,*], Jing Cui [1], Qiang Li [1], Qichao Yao [1], De Zhang [3], Jiachang Guo [1], Xinqian Wang [1] and Chong Qu [2]

[1] National Institute of Natural Hazards, Ministry of Emergency Management of China, Beijing 100085, China; machenyu233@mails.ucas.ac.cn (C.M.); jingcui@ninhm.ac.cn (J.C.); guojiachang241@mails.ucas.ac.cn (J.G.); wangxinqian24@mails.ucas.ac.cn (X.W.)
[2] School of Emergency Management Science and Engineering, University of Chinese Academy of Sciences, Beijing 101408, China; quchong24@mails.ucas.ac.cn
[3] Department of Geography, Faculty of Geographical Science, Beijing Normal University, Beijing 100875, China; 202321051036@mail.bnu.edu.cn
* Correspondence: siquanyang@ninhm.ac.cn

Abstract

Assessing wildfire susceptibility requires integrating environmental and anthropogenic factors to quantify the probability and vulnerability of fires in a given area. Many existing machine-learning models offer high predictive power but limited interpretability, restricting their utility for operational decision-making. This study is the first to apply the intrinsically interpretable deep network TabNet to wildfire susceptibility modeling. By fusing multi-source data and leveraging TabNet's feature-mask matrix, we achieve accurate prediction and built-in explanation without relying on auxiliary tools. On a dataset of 133,811 samples, the proposed model achieves an Area Under the Curve (AUC) of 0.760, recall of 0.883, precision of 0.395, and an F1.5 score of 0.640, outperforming XGBoost (version 1.5.0) and other baseline models. The importance rankings derived from the feature-mask matrix align with the Shapley Additive Explanations (SHAP) results, confirming the reliability of the explanations. This approach combines predictive accuracy with transparency, providing a deployable framework for wildfire early warning, risk management, and ecosystem conservation.

Keywords: wildfire susceptibility; remote sensing; machine learning; interpretable

1. Introduction

Wildfires are uncontrolled fires that burn in wildland vegetation, typically in rural or sparsely populated areas. In recent decades, intensified global warming and escalating human activities have increased the frequency of extreme weather events, making wildfires a major natural hazard that threatens ecosystem security [1]. Beyond immediate damage to natural resources, wildfires can induce long-term and cascading ecological consequences, including biodiversity loss [2], reductions in carbon stocks [3], and elevated atmospheric $PM_{2.5}$ emissions [4], thereby aggravating regional and global air pollution [5]. Globally, the wildfire-burned area in 2024 reached 367 Mha (ranking 17th since 2001), with an estimated 1965 Tg of carbon emitted [6].

Wildfire susceptibility assessment aims to quantify the likelihood of wildfire occurrence across a geographic region by integrating environmental and anthropogenic determinants. Susceptibility maps can support targeted prevention, resource allocation,

and the development of predictive tools for wildfire risk management. Nevertheless, susceptibility assessments remain challenging due to the stochasticity and multidimensional causality of wildfire ignition (e.g., human activities, lightning, and fuel-related spontaneous combustion), which introduce substantial uncertainty. Moreover, wildfire-prone ecosystems are governed by synergistic interactions among thermal conditions, drought intensity, wind, and topography, making single-factor analyses insufficient for comprehensive risk characterization.

To address these complexities, machine learning has been widely adopted for wildfire susceptibility mapping. The core objective is to identify areas that are more likely to be affected by wildfires by learning the relationship between wildfire occurrences and multidimensional environmental drivers. Some studies frame susceptibility mapping as an anomaly detection problem, identifying locations that resemble fire-affected regions [7,8]. More commonly, susceptibility assessment is formulated as a supervised learning task, where classification or regression models are trained using wildfire occurrence labels and environmental features [9–15]. The selection of environmental variables is critical in this setting; commonly used predictors include Land Surface Temperature (LST), the Normalized Difference Vegetation Index (NDVI), the Temperature Vegetation Dryness Index (TVDI), aspect, and wind, which jointly describe fire-prone conditions from multiple perspectives [16]. A variety of algorithms have been explored, including Kernel Logistic Regression (KLR) and Random Forest (RF) [17,18]. Ensemble learning methods, particularly XGBoost and LightGBM, are frequently reported to yield strong performance due to the iterative optimization of weak learners [16,19–21]. More recently, deep learning models such as Artificial Neural Networks (ANN) and U-Net have also been applied and, in some studies, have outperformed ensemble approaches [22–24]. Overall, these efforts demonstrate that data-driven models can substantially enhance susceptibility prediction by capturing complex nonlinear relationships between wildfire occurrence and environmental controls.

Despite these advances, model interpretability remains a key bottleneck for practical deployment. Interpretability refers to the extent to which a model's decision process can be understood, traced, and trusted by humans, thereby enabling the identification of influential drivers and supporting actionable decision-making. Post hoc explanation techniques such as Shapley Additive Explanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME) are commonly used in wildfire susceptibility studies. SHAP, in particular, attributes predictions to individual features based on Shapley values and decomposes outputs into additive contributions [25–27]. However, post hoc explanations can be sensitive to modeling assumptions and baseline selections, and may yield approximate or unstable interpretations, which can limit stakeholder confidence and complicate operational use.

Accordingly, a critical research gap persists in developing wildfire susceptibility models that can simultaneously achieve high predictive accuracy and provide transparent, reliable explanations. To address this gap, we propose a TabNet-based wildfire susceptibility assessment framework. TabNet [28] is a neural network designed for tabular data that integrates a sequential attention mechanism with a differentiable feature-selection module, offering intrinsic interpretability while modeling complex nonlinear relationships. Using Google Earth Engine (GEE), we compile multisource environmental predictors and remote-sensing fire-hotspot observations for the study area from 2010 to 2020, and train an optimal TabNet model using k-fold cross-validation. We evaluate the proposed approach across multiple wildfire events and benchmark it against commonly used machine learning models. In addition to susceptibility mapping, we provide interpretation outputs derived from TabNet's built-in feature selection and attention, enabling the identification of key environmental drivers. In summary, the main contribution of this study is the introduction

https://doi.org/10.3390/fire9030107

and validation of an accurate yet intrinsically interpretable deep learning framework for wildfire susceptibility mapping, which is intended to support wildfire prevention and early-warning applications.

Article Structure Overview: This study applies an intrinsically interpretable deep learning framework (TabNet) to wildfire susceptibility mapping in Southwest China. The article first delineates the study area, describes the data sources, and outlines the preprocessing workflow, in which multi-source climatic, vegetation, topographic, and anthropogenic predictors are integrated with Moderate Resolution Imaging Spectroradiometer(MODIS) FireMask-derived fire occurrence labels on a 1 km spatial grid. It then specifies the TabNet model architecture, sampling strategy, and evaluation protocol, and presents comparative performance analyses against several baseline models. Finally, the article examines model interpretability using TabNet's sparse feature-mask matrices, conducts consistency checks with SHAP-based explanations, discusses the implications and limitations of the approach, and concludes with the principal findings relevant to wildfire risk assessment and management.

2. Materials and Methods

2.1. Study Area

The study area is located in southwestern China (21°08′–34°19′ N, 97°21′–110°11′ E), covering approximately 962,500 km^2 (Figure 1). The region contains extensive forest resources and serves as an important ecological barrier for the upper Yangtze River Basin(China) [29]. It is dominated by a subtropical plateau monsoon climate with distinct dry and wet seasons [30]. The dry season (November–April) features low precipitation, strong winds, prolonged sunshine, and low fuel moisture, resulting in elevated fire danger [31]. In contrast, the wet season (May–October) is characterized by high humidity and frequent rainfall.

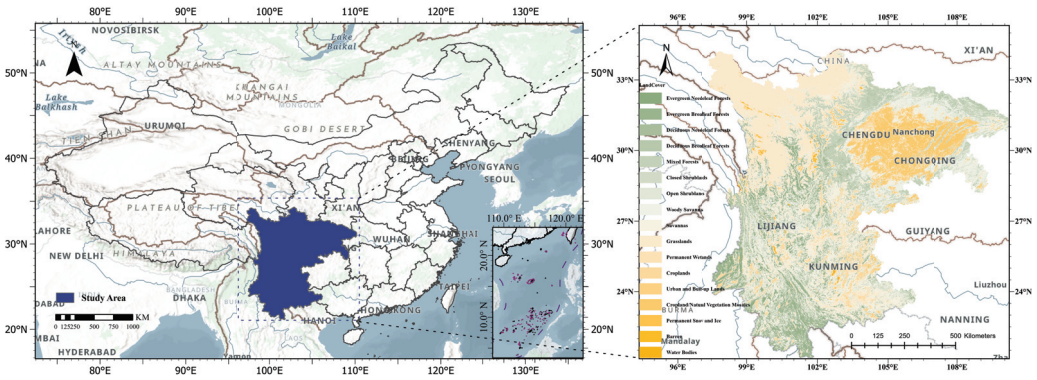


Figure 1. Location of the study area in Southwestern China. The (**left**) figure shows the location of the study area, with the study region highlighted in blue. The land-cover map (**right**) was derived from the MODIS Land Cover Type product (MOD11A2) following the IGBP classification scheme. Major geographical features and the boundary of the primary study area are annotated.

The rugged terrain—high mountains, deep valleys, and steep slopes—facilitates wildfire occurrence, especially in mountainous areas at elevations of 2000–3500 m [32]. Seasonal hydroclimatic contrasts and the complex topography jointly increase regional wildfire susceptibility, threatening livelihoods, infrastructure, and ecosystem stability [33].

Wildfires increasingly affect wildland–urban interfaces, posing risks to critical infrastructure and biodiversity hotspots [34]. Fire suppression is further constrained by hypoxic high-elevation conditions, which can substantially reduce firefighters' physical work capacity compared with lowland environments. Consequently, the region remains one of the most challenging areas in China for wildfire prevention and control. These conditions mo-

https://doi.org/10.3390/fire9030107

tivate the development of high-precision and interpretable wildfire susceptibility models to reduce escalating ecological and socioeconomic losses [35].

2.2. Data and Data Preprocessing

This study incorporated both natural and anthropogenic determinants directly associated with forest fire occurrence for modeling analyses. The natural determinants encompassed climatic, vegetative, and topographic attributes, whereas anthropogenic determinants primarily reflected the presence of ignition sources and the potential for fire spread. Fire mask data were obtained from the MODIS (MOD14A1) thermal anomaly product. Climatic variables included air temperature, precipitation, wind speed, wind direction, and soil moisture. Vegetation-related parameters comprised the Normalized Difference Vegetation Index (NDVI), the 3-month Standardized Precipitation Evapotranspiration Index (SPEI3), the Normalized Difference Infrared Index 7 (NDII7) [36], and land cover categories. Topographic characteristics were represented by elevation, and anthropogenic influence was proxied by population density. In addition, the day-of-year corresponding to the fire occurrence time was incorporated to characterize seasonal variability in fire activity. Detailed parameter descriptions and associated metadata are presented in Table 1.

Table 1. Detailed description of data sources.

Data Type	Description	Data Source
Soil Moisture	Monthly average soil moisture at 0–10 cm depth	NASA/FLDAS/NOAH01/C/GL/M/V001
SPEI_03	3-month Standardized Precipitation Evapotranspiration Index (SPEI)	CSIC/SPEI/2–9
FireMask	Binary mask indicating fire points	MODIS/061/MOD14A1
NDII7	Short-wave infrared moisture index, calculated from MODIS reflectance bands	MODIS/006/MOD09GA
Precipitation	Daily total precipitation in mm	NASA/GPM_L3/IMERG_V07
LST	Daily land surface temperature in Kelvin	MODIS/061/MOD11A1
Wind Speed	Daily maximum wind speed at 10 m height in m/s	ECMWF/ERA5/DAILY
Wind Direction	Wind direction at 10 m height in degrees	ECMWF/ERA5/DAILY
Air Temperature	2 m daily mean air temperature in Celsius	ECMWF/ERA5/DAILY
Elevation	30 m digital elevation model in meters	USGS/SRTMGL1_003
NDVI	16-day composite vegetation index	MODIS/061/MOD13A1
Land Cover	IGBP classification of land cover types	MODIS/061/MCD12Q1
Population Density	Gridded population density (persons per unit area)	CIESIN/GPWv411/GPW_Population_Density
Day of the Year	Fire occurrence day in the year	MODIS/061/MCD64A1

All data aggregation and preprocessing were conducted on the Google Earth Engine (GEE) platform. Prior to model development, preprocessing comprised two primary tasks: (1) the acquisition and georeferencing of raw data and (2) the sampling and assembly of the modeling dataset. Among the candidate features, the Normalized Difference Infrared Index (NDII) was computed using the standard formulation (Equation (1)) and served as a key fire-related variable. In this study, NDII7 denotes the NDII, computed using the MODIS band 7 reflectance as the shortwave-infrared (SWIR) component, hence the suffix "7". Climate variables (precipitation, temperature, and wind speed) were imported directly from their respective datasets without further processing.

The modeling dataset was constructed in three steps. First, all raster layers covering the study area were resampled to a uniform 1 km $\times$ 1 km grid to ensure consistent spatial resolution. Second, samples were generated using the MOD14A1 FireMask layer, whose pixel classes explicitly distinguish fire, non-fire, and no-observation conditions. Specifically, pixels with FireMask values 7–9 were labeled as positive samples (fire detections with confidence levels), while pixels with FireMask value 5 (non-fire land pixel) were labeled as negative samples. Pixels flagged as not processed, water, cloud, or unknown (e.g., FireMask values 1–4 and 6) were excluded from negative sampling to avoid introducing label noise

https://doi.org/10.3390/fire9030107

from unobserved/invalid conditions. This choice follows the official FireMask class definitions and ensures that the negative class represents valid non-fire land observations [37]. To reduce spatial dependence and potential near-fire label contamination, we applied a 5 km buffer around historical fire detections. Positive samples were randomly selected within the buffered regions, while negative samples were drawn from areas outside all buffers. The 5 km radius represents an operational trade-off given the 1 km spatial resolution of the harmonized predictors and the MOD14A1 fire mask, and is consistent with the distance-based separation practices commonly used to mitigate spatial autocorrelation effects in spatial prediction. Third, because the focus of this study is forest fire occurrence, we applied an IGBP-based land-cover mask to restrict sampling to forest-dominated fuel environments and excluded non-target or essentially non-burnable classes (e.g., urban areas, water bodies, barren/sparsely vegetated surfaces, and non-forest vegetation types such as grasslands and wetlands). This masking reduces the risk of trivially separable negatives (e.g., open water) inflating model performance and keeps model interpretation centered on forest-fuel fire drivers. We note that the resulting model is intended for forested landscapes and should not be directly interpreted as a universal fire model across all land-cover regimes.

Missing values and outliers were identified and treated using the Isolation Forest algorithm to improve data quality.

Following these procedures, we obtained a structured binary wildfire dataset—labeled by FireMask values—suitable for model training. The final dataset comprises $N = 133{,}811$ samples with $d = 13$ features, as shown in Figure 2.

$$\mathrm{NDII7} = \frac{\rho_{\text{Near Infrared}} - \rho_{\text{SWIR (MODIS Band 7)}}}{\rho_{\text{Near Infrared}} + \rho_{\text{SWIR (MODIS Band 7)}}} \tag{1}$$

where:

- $\rho_{\text{Near Infrared}}$ is the reflectance in the near-infrared band (dimensionless);
- $\rho_{\text{SWIR (MODIS Band 7)}}$ is the reflectance in the shortwave infrared band from MODIS band 7 (dimensionless).

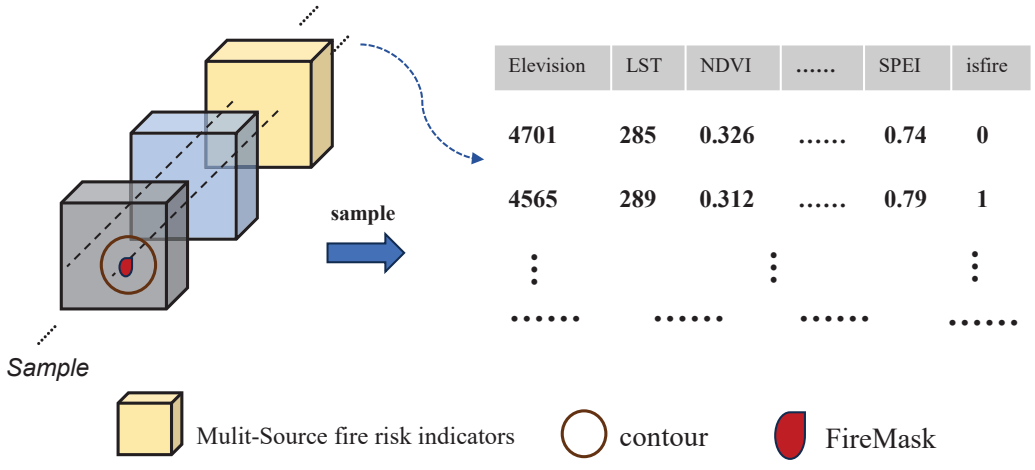


Figure 2. Schematicdiagram of the sampling process used to construct the modeling dataset from multi-source inputs. Red polygons denote positive-sample regions, red circles indicate the 5 km buffer zones around historical fire points, and negative samples are drawn from areas outside the buffers.

Following the aforementioned procedure, the heterogeneous environmental parameters obtained from multiple sources were converted into labeled, structured datasets suitable for subsequent input into the model.

https://doi.org/10.3390/fire9030107

Reproducibility and Sensitivity to Key Assumptions

All data aggregation, preprocessing, and sampling were implemented on the Google Earth Engine (GEE) platform with explicitly recorded dataset identifiers (collection/version IDs), spatial harmonization rules (continuous vs. categorical resampling), FireMask labeling and exclusion rules, buffer radius, random seed(s), and outlier screening settings. These key parameters are summarized in Table 2 to facilitate the exact reproduction of the modeling dataset.

Table 2. Key preprocessing, sampling, and reproducibility parameters (GEE workflow).

Component	Setting/Rule	Value/Notes
Study period	Temporal coverage of all variables and labels	January 2010 to September 2020
Spatial grid	Target spatial resolution and grid definition	1 km × 1 km; CRS: EPSG:4326 (WGS84)
Resampling (continuous)	Resampling/aggregation method for continuous predictors	Nearest-neighbor aggregation/resampling on GEE (applied to continuous predictors after harmonization to 1 km)
Resampling (categorical)	Resampling method for categorical predictors (land cover)	As described in Section 2.2 (IGBP land-cover classes preserved during 1 km harmonization); exact categorical resampling operator not separately reported
Fire label definition	Positive and negative label mapping from MOD14A1 FireMask	Positive (fire): FireMask = 7, 8, 9; Negative (non-fire): FireMask = 5
Fire label exclusions	Excluded FireMask categories (invalid/unobserved conditions)	Not explicitly reported in the main text; see the released dataset documentation and sampling script/configuration in the GitHub repository https://github.com/machenyu2023/wildfire-dataset (accessed on 28 December 2025)
Sampling design	Spatial sampling rule for positives and negatives	Positive samples randomly selected within buffers around historical fire detections; negative samples randomly drawn from areas outside all buffers
Buffer radius	Radius for spatial separation around historical fire detections	5 km
Sampling balance	Class balance and sampling size control	Final dataset size: $N = 133{,}811$; class ratio/stratification scheme not separately reported (see repository documentation)
Randomness control	Random seed(s) for sampling and preprocessing	Random seed(s) not separately reported in the manuscript; see repository documentation
Missing value handling	Masking/imputation rule prior to modeling	Not separately reported; missing values and outliers were handled during data cleaning (see Section 2.2 and repository documentation)
Outlier screening	Outlier detection algorithm and hyperparameters	Isolation Forest used for outlier identification; hyperparameters not separately reported (see repository documentation)
Outlier impact	Fraction of samples removed as outliers (optional but recommended)	Not separately reported
Data versioning	Dataset identifiers (GEE collection/version IDs)	All input datasets and GEE collection/version IDs are listed in Table 1 (Data Sources)

To support reproducibility, we have released the processed wildfire dataset (including labels and predictors) and accompanying documentation on GitHub at (https://github.com/machenyu2023/wildfire-dataset (accessed on 28 December 2025)). This repository enables readers to reconstruct the modeling inputs and to conduct straightfor-

https://doi.org/10.3390/fire9030107

ward sensitivity checks of key assumptions (e.g., alternative buffer radii or land-cover masks) using the same workflow.

2.3. Model Selection and Methodology

This study utilizes TabNet for a wildfire susceptibility assessment. The TabNet architecture is shown in Figure 3, where the encoder–decoder-based framework consists of sequential decision steps. Each decision step incorporates two key components: the Attentive Transformer and the Feature Transformer. The Attentive Transformer dynamically selects key features using a sparse feature mask, while the Feature Transformer processes features through shared and independent layers to extract global and local patterns. The decoder aggregates contributions from all steps and employs residual connections to optimize gradient propagation, ensuring efficient end-to-end learning.

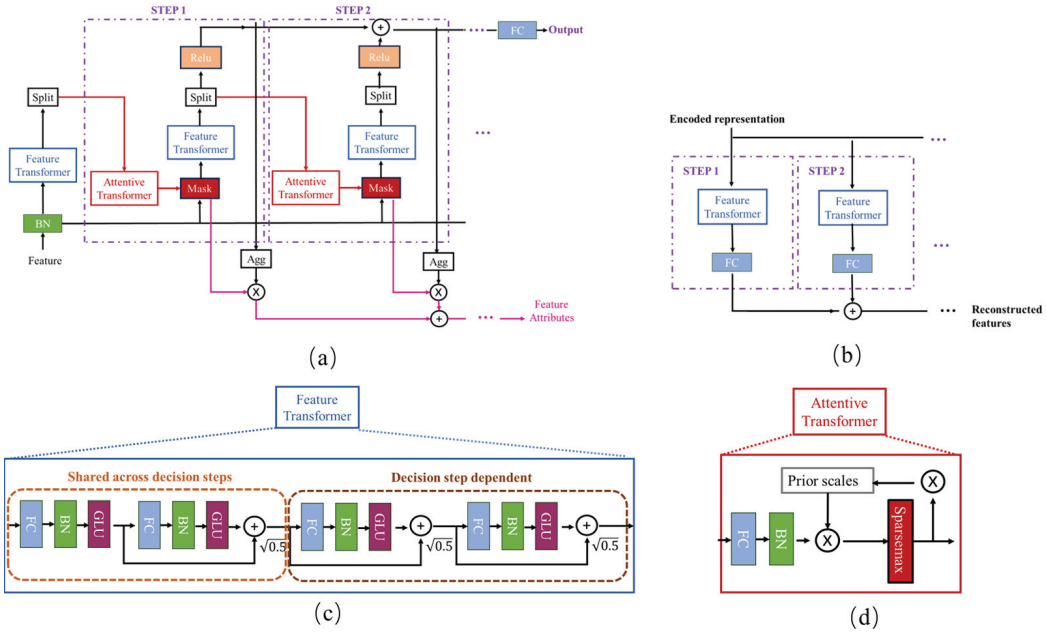


Figure 3. TabNet architecture: (**a**) Encoder, (**b**) Decoder, (**c**) Feature Transformer, (**d**) Attentive Transformer.

To address class imbalance in the wildfire dataset, we apply three strategies: class weighting, SMOTE oversampling, and random undersampling. Class weighting adjusts the loss function to amplify minority class errors. SMOTE generates synthetic minority class samples, enriching decision boundaries and reducing overfitting. Random undersampling reduces majority class samples, balancing class representation.

Following feature standardization via Batch Normalization, features are processed through the TabNet encoder. Each decision step computes attention weights using the previous step's feature vector and scale information, followed by sparse feature masking via the Sparsemax function. The selected features undergo transformation through GLU-based feature transformers, producing decision vectors that are aggregated to form the final output.

Model training minimizes a weighted binary cross-entropy loss to account for class imbalance:

$$L_{\mathrm{CE}} = -\sum_{i} w_i [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \tag{2}$$

where w_i denotes class weights, and y_i and $\hat{y}_i$ represent the true labels and predicted probabilities, respectively. We use the Adam optimizer with learning rate scheduling and

https://doi.org/10.3390/fire9030107

early stopping. Hyperparameter optimization is conducted using Bayesian optimization (TPE) to maximize the F_β score:

$$F_\beta = (1+\beta^2)\frac{\text{Precision} \times \text{Recall}}{\beta^2 \cdot \text{Precision} + \text{Recall}} \tag{3}$$

This process fine-tunes the hyperparameters: n_d, n_a, n_{steps}, γ, and the classification threshold.

TabNet's interpretability is facilitated by its sparse feature masking, providing both local and global insights. Local interpretability is achieved by examining the sparse feature masks at each step, revealing the features used for decision-making at each step. Global interpretability is derived by aggregating these masks across all steps, weighted by the magnitude of the decision vectors:

$$M_{\text{agg}} = \sum_{i=1}^{N_{\text{steps}}} \eta_i M_i \tag{4}$$

Sparse feature selection is encouraged by a regularization term that enforces sparsity on the masks:

$$L_{\text{sparse}} = \sum_{i=1}^{N_{\text{steps}}} \sum_{b=1}^{B} \sum_{j=1}^{D} -M_{i,b,j} \log(M_{i,b,j} + \varepsilon) \tag{5}$$

where ε ensures numerical stability. Visualization of the aggregated global importance and individual step masks enables better understanding of the model's decision pathways.

2.4. Model Evaluation and Comparison

Model performance is evaluated using standard classification metrics—Precision, Recall, and the F_β-score—each characterizing a distinct dimension of predictive capability. These metrics are defined as

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \qquad \text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{6}$$

$$F_\beta\text{-score} = \frac{(1+\beta^2) \cdot \text{Precision} \cdot \text{Recall}}{\beta^2 \cdot \text{Precision} + \text{Recall}} \tag{7}$$

where TP, FP, TN, and FN denote True Positives, False Positives, True Negatives, and False Negatives, respectively.

In the context of wildfire susceptibility assessment, the cost of false negatives (i.e., missed ignitions) typically exceeds that of false positives, as a single undetected fire can rapidly escalate into a large-scale incident, whereas an occasional false alarm primarily results in the unnecessary deployment of monitoring or suppression resources. The F_β-score explicitly encodes this asymmetry by weighting Recall β times more strongly than Precision; for $\beta = 1$, both contribute equally. In this study, we set $\beta = 1.5$, effectively assigning a 50% higher weight to Recall relative to Precision. This choice favours models that minimise missed fire events, even at the cost of a moderate increase in false alarms, aligning the evaluation criterion with the operational preferences of wildfire management agencies.

In this study, a hierarchical random sampling strategy was employed to partition the original dataset into three subsets: a training set (60%), a validation set (20%), and an independent test set (20%). The process involved the following steps:

- Independent test set retention : 20% of the original samples were set aside as an independent test set, with fixed random seeds to ensure reproducibility.

https://doi.org/10.3390/fire9030107

- Category-balanced processing: The remaining 80% of the data underwent category-balancing treatment, which included optional oversampling (SMOTE) or undersampling techniques to address class imbalance.
- Stratified sampling: The balanced set was then divided into a training subset (75%) and a validation subset (25%) using stratified sampling, ensuring consistent class distribution across the subsets and minimizing sampling bias.

This careful division of data helps ensure the model's robustness, reduces potential biases, and improves the generalizability of results.

For model training, a three-stage optimization framework was applied:

- Hyperparameter search: Hyperparameters were optimized using Bayesian optimization (Hyperopt TPE) combined with K-fold hierarchical cross-validation. In each iteration, the features were standardized using Z-score normalization, and the mean $F_{1.5}$ score across validation folds was used as the optimization objective.
- Retraining with complete training data: The model was then retrained using the entire training set, including both training and validation data. The optimal classification threshold was determined via grid search with a step size of 0.01.
- Final evaluation: The final model performance was assessed on the independent test set using metrics such as the $F_{1.5}$ score, precision, recall, and AUC.

To ensure reliability, all random processes used fixed seeds, and early stopping (with a patience of 10 epochs) was implemented to prevent overfitting. A total of 30 optimization rounds with 10-fold cross-validation were performed to fine-tune the model and assess its performance.

Additional diagnostic information is obtained from the Receiver Operating Characteristic (ROC) curve, which plots the True Positive Rate (TPR) against the False Positive Rate (FPR) across varying decision thresholds. The Area Under the Curve (AUC) provides a scalar summary of overall discriminative ability, with values ranging from 0.5 (no skill, equivalent to random guessing) to 1.0 (perfect discrimination). Together, these metrics afford a multidimensional characterization of model robustness under varying class distributions and decision thresholds.

Model validation is performed using spatiotemporal wildfire event datasets sourced from Google Earth Engine (GEE), which integrate environmental covariates with validated FireMasks. A multi-event evaluation protocol is employed to rigorously assess the model's generalization capability, thereby increasing confidence in its applicability for operational wildfire risk forecasting. For the case study validation, a decision threshold of 0.6 was selected, where regions with a fire likelihood greater than 0.6 are classified as high-risk (fire occurrence). This threshold was chosen to balance the model's ability to identify high-risk areas while minimizing both false positives and false negatives. A higher threshold, such as 0.7 or above, could result in too many areas being incorrectly categorized as low-risk, whereas a lower threshold (e.g., 0.5 or below) would increase the occurrence of false positives. The threshold of 0.6 thus provides a suitable compromise. In creating the wildfire susceptibility maps, kriging interpolation and natural breaks classification were employed for visualization.

For comparative analysis, three baseline models were considered:

1. Random Forest (RF): A classical ensemble learning algorithm based on bootstrap aggregation of decision trees.
2. XGBoost: A gradient boosting framework employing second-order optimization of the loss function.
3. LightGBM: A high-efficiency gradient boosting implementation using leaf-wise tree growth with histogram-based splitting.

https://doi.org/10.3390/fire9030107

All models were trained using an identical preprocessing pipeline and evaluated under a five-fold cross-validation scheme to ensure a fair and statistically robust comparison.

3. Results and Case Studies

3.1. Results Overview

This section reports (i) the training and evaluation results of the proposed TabNet model, (ii) spatial validation using three historical wildfire incidents, and (iii) comparisons against benchmark machine-learning models.

3.2. Model Training Results

As summarized in Table 3, the final TabNet configuration was selected from a predefined search space, including $n_d = n_a = 16$, $n_{\text{steps}} = 6$, $\gamma = 1.0$, $n_{\text{independent}} = 1$, $n_{\text{shared}} = 2$, $\lambda_{\text{sparse}} = 0.001$, and a learning rate of 0.005. Using this configuration, the model achieved strong validation performance (Table 4) with $F_{1.5} = 0.9701$, precision of 0.9274, and recall of 0.9903 on 48,433 samples with 13 features. On the independent test set (26,763 samples; 13 features), the performance was $F_{1.5} = 0.7746$, with precision of 0.5416 and recall of 0.9577.

Table 3. TabNet hyperparameters, including their definitions, search spaces, and the selected values after tuning for optimal model performance.

Hyperparameter	Definition	Search Space	Selected
n_d	Decision layer dimension in feature processing	{8, 16, 24}	16
n_a	Attention layer width for feature selection	{8, 16, 24}	16
n_{steps}	Number of sequential processing steps	{4, 5, 6}	6
γ	Mask relaxation coefficient (sparsity control)	{1.0, 1.2}	1.0
$n_{\text{independent}}$	Independent GLU blocks per step	{1, 2}	1
n_{shared}	Shared GLU blocks across steps	{2, 3}	2
λ_{sparse}	Regularization for sparse feature selection	{0.001, 0.005, 0.01}	0.001
Learning rate	Learning rate configuration	{0.002, 0.005}	0.005

Table 4. Model evaluation results on validation and independent test datasets, reporting performance metrics including $F_{1.5}$ score, precision, and recall for wildfire detection.

Dataset	Samples (n)	Features (d)	$F_{1.5}$	Precision	Recall
Validation	48,433	13	0.9701	0.9274	0.9903
Test	26,763	13	0.7746	0.5416	0.9577

Note: $F_{1.5}$ emphasizes recall over precision, reflecting the higher cost of missing wildfire detections in practical applications.

3.3. Case Study Results

This section presents the evaluation of several significant wildfire events through the application of a model optimized to assess environmental features on the respective event days. These features were derived from historical data and spatially matched against fire-point observations from the MODIS (MOD14A1) satellite. The primary objective was to investigate the correlation between model outputs and actual wildfire occurrences. The resulting data were visualized using ordinary kriging interpolation to generate wildfire susceptibility maps. While five case studies were initially selected, the analysis focuses on three cases for visual representation.

Table 5 summarizes the fire-point prediction accuracy of the selected cases (where samples with a confidence greater than 0.6 are defined as fire points, see Section 2.4), showing the number of MODIS fire points, the matched points, and the matching rate.

https://doi.org/10.3390/fire9030107

Figures 4–6 display comparisons between optical imagery and model-derived susceptibility for the Muli, Kangding, and Lijiang wildfire events. These figures provide insight into the spatial distribution of fire risk within the regions, with high-risk areas often aligning with observed fire points.

Table 5. Fire-point prediction accuracy comparison between model predictions and MODIS (MOD14A1) observations, showing the number of MODIS fire points, the number of matched points, and the matching rate for different locations and dates.

Location	Date	MODIS Fire Points (n)	Matched (n)	Rate (%)
XiChang	30 March 2020	41	13	31.7
MuLi	16 January 2017	28	0	0.0
MuLi	30 March 2020	109	86	78.9
LiJiang	7 February 2017	89	45	50.6
KangDing	28 February 2013	45	20	44.4

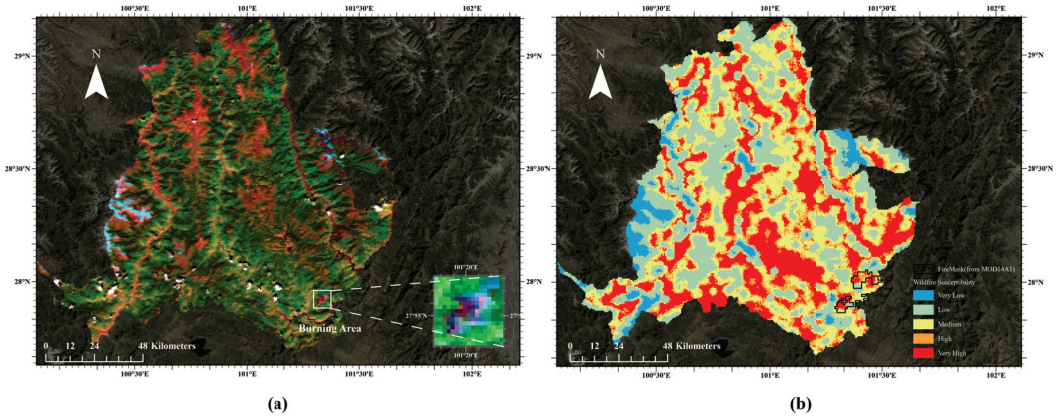


Figure 4. Comparison of optical imagery and model-derived susceptibility for the Muli wildfire event. (**a**) MODIS false-color composite (SWIR–Green–Blue) over the Muli region, acquired on 30 March 2020; the white overlay highlights the burning area. (**b**) Wildfire susceptibility map from model inference, based on MODIS data acquired on 30 March 2020; polygons denote FireMask vectors extracted from MOD14A1.

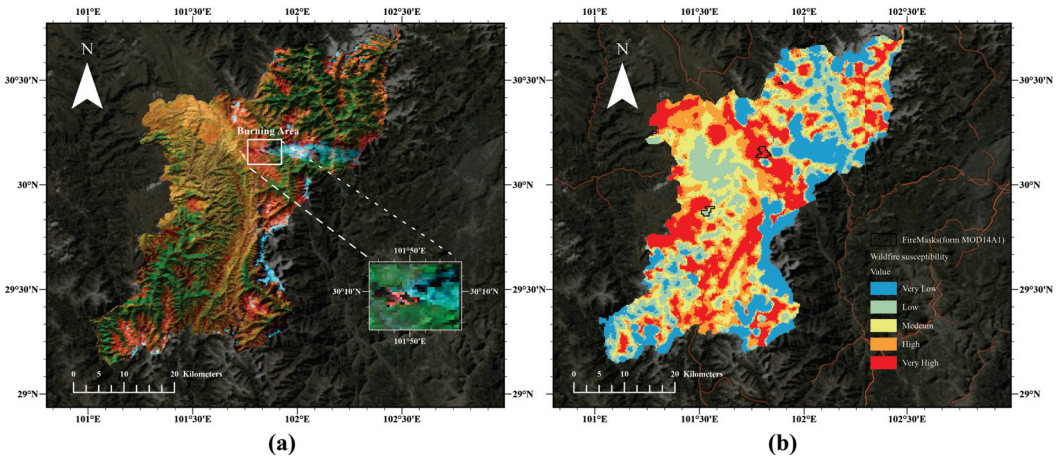


Figure 5. Comparison of optical imagery and model-derived susceptibility for the Kangding wildfire event. (**a**) MODIS false-color composite (SWIR–Green–Blue) acquired on 28 February 2013 over the Kangding region; the white overlay highlights the burning area. (**b**) Wildfire susceptibility map from model inference; polygons denote FireMask vectors extracted from MOD14A1.

https://doi.org/10.3390/fire9030107

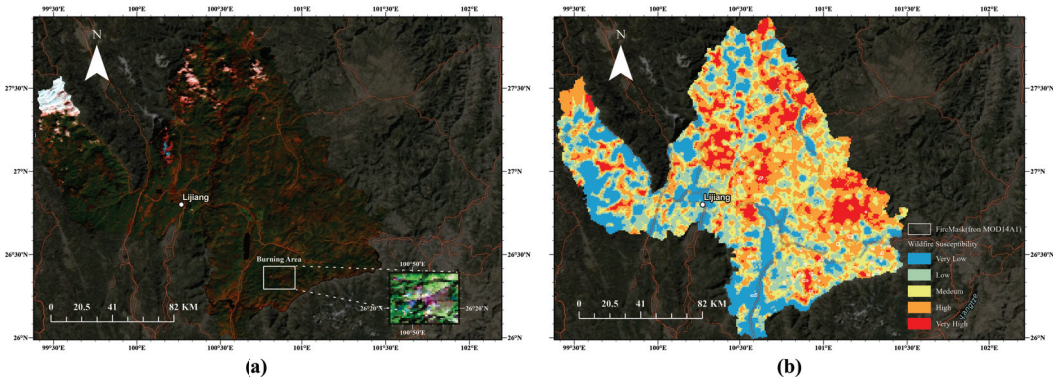


Figure 6. Comparison of optical imagery and model-derived susceptibility for the Lijiang wildfire event. (**a**) MODIS false-color composite (SWIR–Green–Blue) acquired on 7 February 2017 over the Lijiang region; the white overlay highlights the burning area. (**b**) Wildfire susceptibility map from model inference; polygons denote FireMask vectors extracted from MOD14A1.

3.4. Model Comparison

Table 6 shows that TabNet produces relatively fewer positive predictions while maintaining a non-trivial proportion of high-confidence predictions, which benefits map credibility. Figure 7 shows that TabNet achieves the highest AUC (0.76), indicating better class separability. Figure 8 suggests that Random Forest produces overly homogeneous regions, while gradient-boosting baselines show limited recognition of adjacent similar environments; TabNet yields more coherent susceptibility gradients in complex geographical settings.

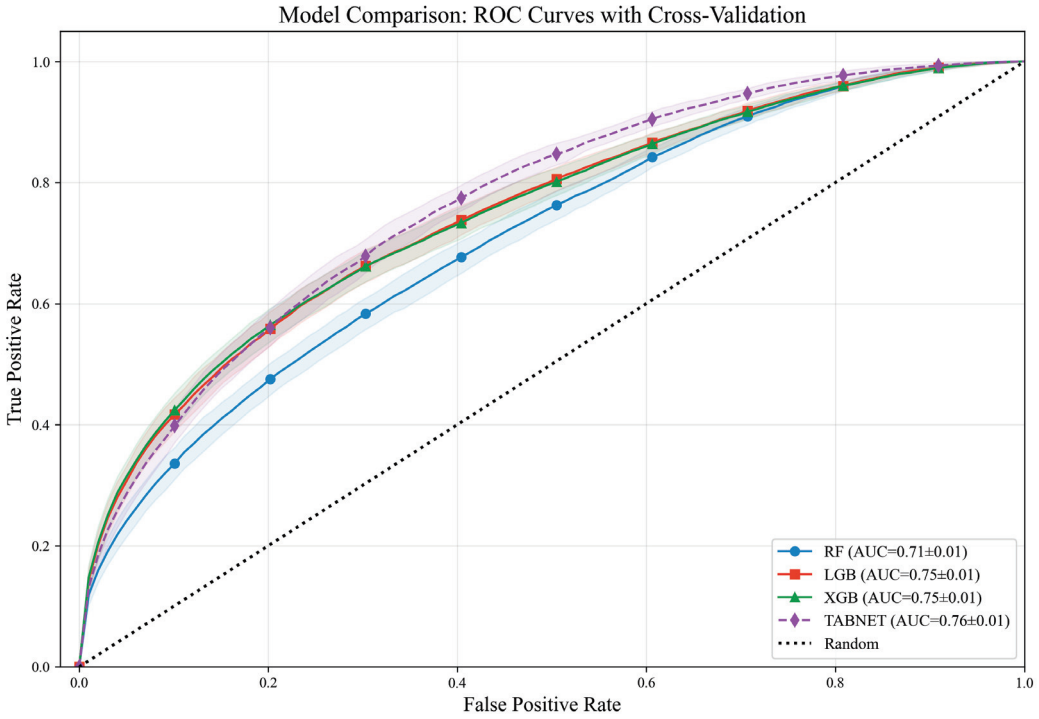


Figure 7. Receiver Operating Characteristic (ROC) curves for the comparison of model performance across multiple evaluation metrics, highlighting the trade-off between sensitivity and specificity.

https://doi.org/10.3390/fire9030107

Table 6. Performance comparison of forest fire prediction models.

Region	Samples	Random Forest			LightGBM			XGBoost			TabNet		
		Fires	>70%	>90%	Fires	>70%	>90%	Fires	>70%	>90%	Fires	>70%	>90%
Xichang	2434	1814	74.5	0.0	1510	62.0	0.2	1397	12.6	0.1	**807**	**13.8**	**2.9**
Kangding	13,355	8508	63.7	0.0	4197	5.7	0.5	3676	6.8	0.3	**4443**	**19.5**	**4.5**
Muli (2017)	15,000	205	1.4	0.0	395	1.0	0.2	839	1.0	0.2	**456**	**1.2**	**0.3**

Note: "Fires (n)" denotes the number of samples predicted as fire. "Conf > 70%" and "Conf > 90%" denote the percentage of predictions whose confidence exceeds the given threshold. The bolded numbers represent the results of this study.

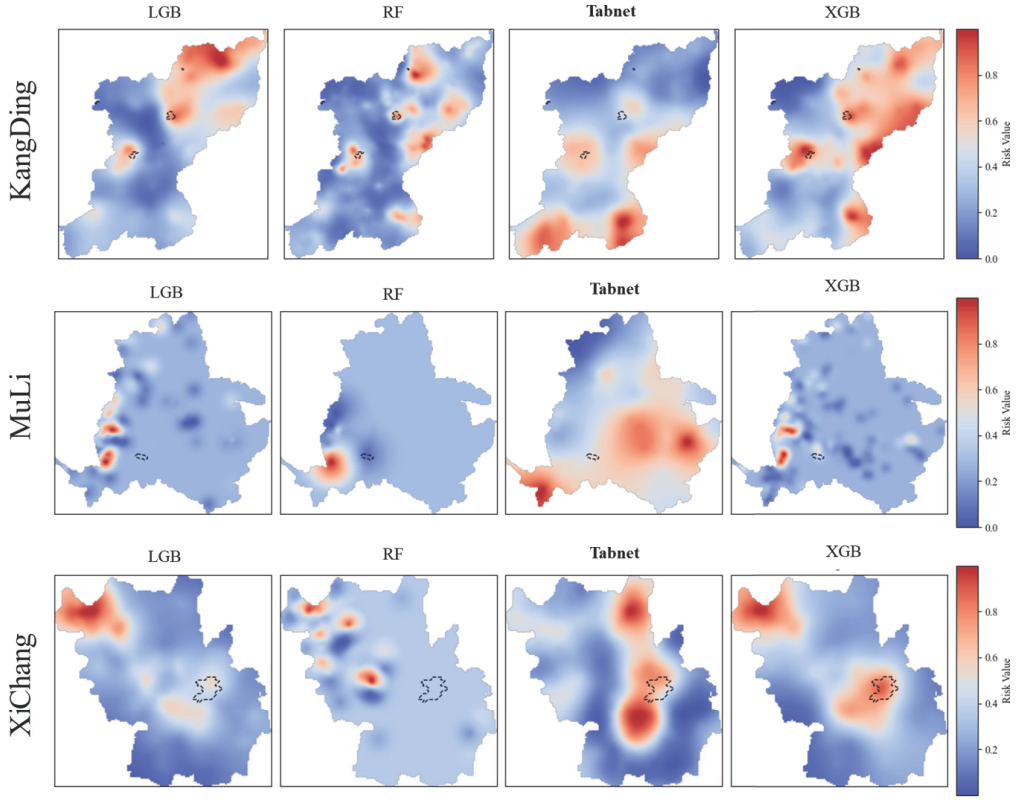


Figure 8. Comparison of wildfire susceptibility maps generated by different machine learning models (**left** to **right**: LightGBM, Random Forest, TabNet, and XGBoost), showcasing the spatial variations in predicted susceptibility across models.

The evaluation of these three case studies reveals that the model generally provides good spatial generalization, with fire points frequently located within high-risk zones. However, the Muli (2017) case, despite a large dataset, shows no matches (0%) with MODIS fire points. This discrepancy warrants further exploration in the Discussion section, as it suggests the influence of additional factors not accounted for in the model's current features.

4. Model Interpretability and Discussion

4.1. Model Performance and Weakness

By comparing precision and recall and calculating $F_{1.5}$, we evaluated the suitability of different models for wildfire susceptibility analysis. As shown in Figure 9, TabNet outperforms the other models in both precision and recall. Compared with XGBoost, TabNet improves precision from 0.3661 to 0.395 (+7.89%) and recall from 0.8675 to 0.8835 (+1.84%). Compared with LightGBM, precision improves by 12.36% (0.3515 → 0.395) and recall by 2.49% (0.8620 → 0.8835). Against Random Forest, precision nearly doubles (0.2001 → 0.395) and recall increases by 24.45% (0.7098 → 0.8835). Since, in many real-

https://doi.org/10.3390/fire9030107

world scenarios, missed detections can be more costly than false alarms, $F_{1.5}$ is often more aligned with practical needs, and TabNet performs well on this metric.

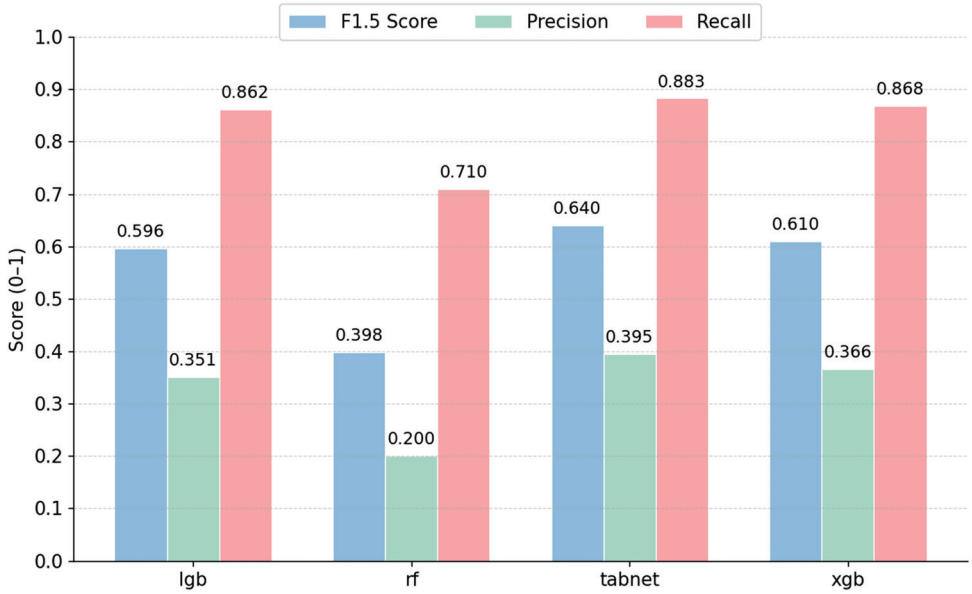


Figure 9. Performance comparison across models using precision, recall, and $F_{1.5}$.

TabNet's advantages are related to its "Feature Transformer–Sparse Attention–Stepwise Decision" architecture, which can suppress overfitting on noisy and imbalanced structured data while retaining an interpretable feature-selection chain. For example, Figure 8 indicates that TabNet assigns the Muli ground–truth area to a higher–susceptibility zone even when direct matching is poor, which may imply improved spatial generalization in complex environments.

4.1.1. Weakness Discussion Based on the Muli Case

The Muli incident also reveals a practical weakness of the alternative baselines: for this event, the high–susceptibility zones produced by XGBoost, LightGBM, and Random Forest do not provide effective spatial matching with the ground–truth burned area (i.e., most of the affected region is not covered by their predicted high-risk classes). This "non–matching" behavior may stem from their reliance on relatively global, non-stepwise feature interactions and split–based decision boundaries, which can be sensitive to local domain shifts (e.g., micro–topography, heterogeneous vegetation mosaics, or strong spatial autocorrelation). In addition, if the Muli area contains atypical combinations of drivers (e.g., elevation–slope–fuel–meteorology patterns) that are under–represented in the training samples, tree-ensemble baselines may regress toward more common patterns and thus fail to highlight the event footprint. Therefore, beyond aggregate metrics, the Muli case emphasizes the necessity of event–level evaluation: a model can achieve reasonable overall recall while still missing critical extreme events in specific geomorphic or climatic contexts.

4.1.2. Limitations and Stability Considerations

A key limitation of interpretability analyses based on feature masks (or SHAP values) is the potential instability of feature rankings under different data splits, random seeds, or parameter settings. In practice, wildfire data may exhibit spatial autocorrelation and class imbalance, which can further amplify variance in estimations.

https://doi.org/10.3390/fire9030107

4.2. Feature Importance and Geographical Distribution

The TabNet model offers intrinsic interpretability by producing feature-mask matrices during inference. As shown in Figure 10, the horizontal axis corresponds to the input features and the vertical axis corresponds to samples; each sub-panel represents a decision step. Brighter values indicate that the corresponding feature is more heavily used at that step. Aggregating these masks yields a global ranking of feature importance (Figure 11). Overall, thermal-related variables (e.g., LST) and vegetation-related factors (e.g., LandCover and NDVI) contribute most to the susceptibility prediction, followed by seasonal timing (day_of_year). Meteorological and moisture-related variables (e.g., precipitation, soil moisture, wind) provide secondary but non-negligible support, suggesting that TabNet learns a driver hierarchy that aligns with wildfire-prone conditions driven by heat, fuel availability, and seasonality.

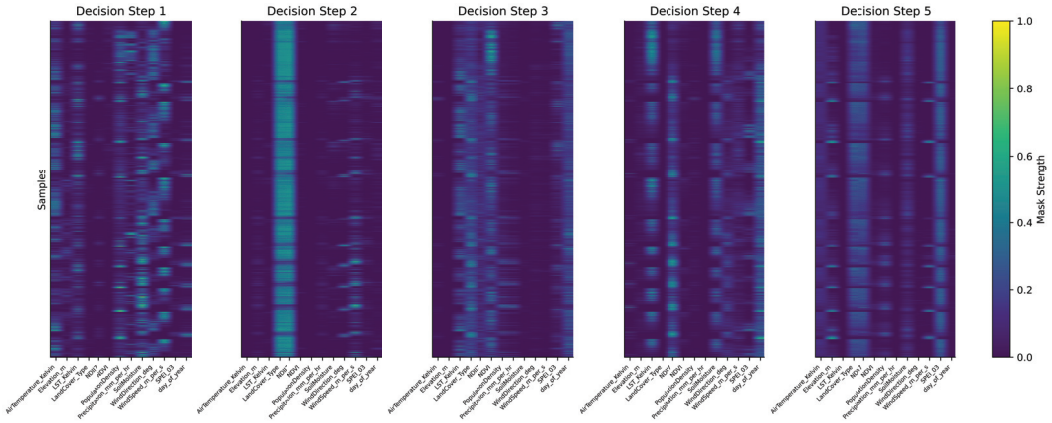


Figure 10. Mask visualizations for wildfire datasets (133,381 samples). Brighter bands indicate greater involvement of the corresponding feature in classification at each decision step. The x-axis labels for each step, from left to right, are: AirTemperature_Kelvin, Elevation_m, LST_Kelvin, LandCover_Type, NDVI, PopulationDensity, Precipitation_mm_per_hr, SoilMoisture, WindDirection_deg, WindSpeed_m_per_s, SPEI_03, and day_of_year.

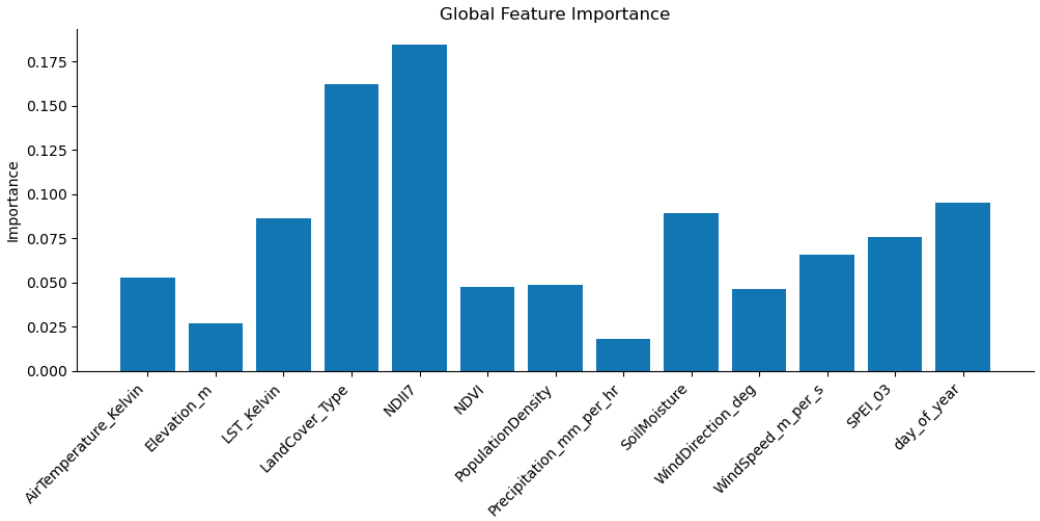


Figure 11. Global feature importance derived from TabNet masks.

We further evaluate event-level interpretability to reveal geographic heterogeneity in the learned drivers. Figure 12 visualizes the masks for three representative incidents, while Figure 13 summarizes the corresponding feature-importance patterns. Although the incidents share a broadly consistent structure (i.e., thermal and vegetation-related variables remain influential), the relative emphasis on specific features varies across events, reflecting

https://doi.org/10.3390/fire9030107

differences in local climate, topography, and human activity. For example, mountainous and strongly seasonal regions tend to show stronger coupling between elevation/seasonal indicators and vegetation-related signals, whereas other events exhibit comparatively higher reliance on moisture or wind-related variables, consistent with event-dependent ignition and spread mechanisms.

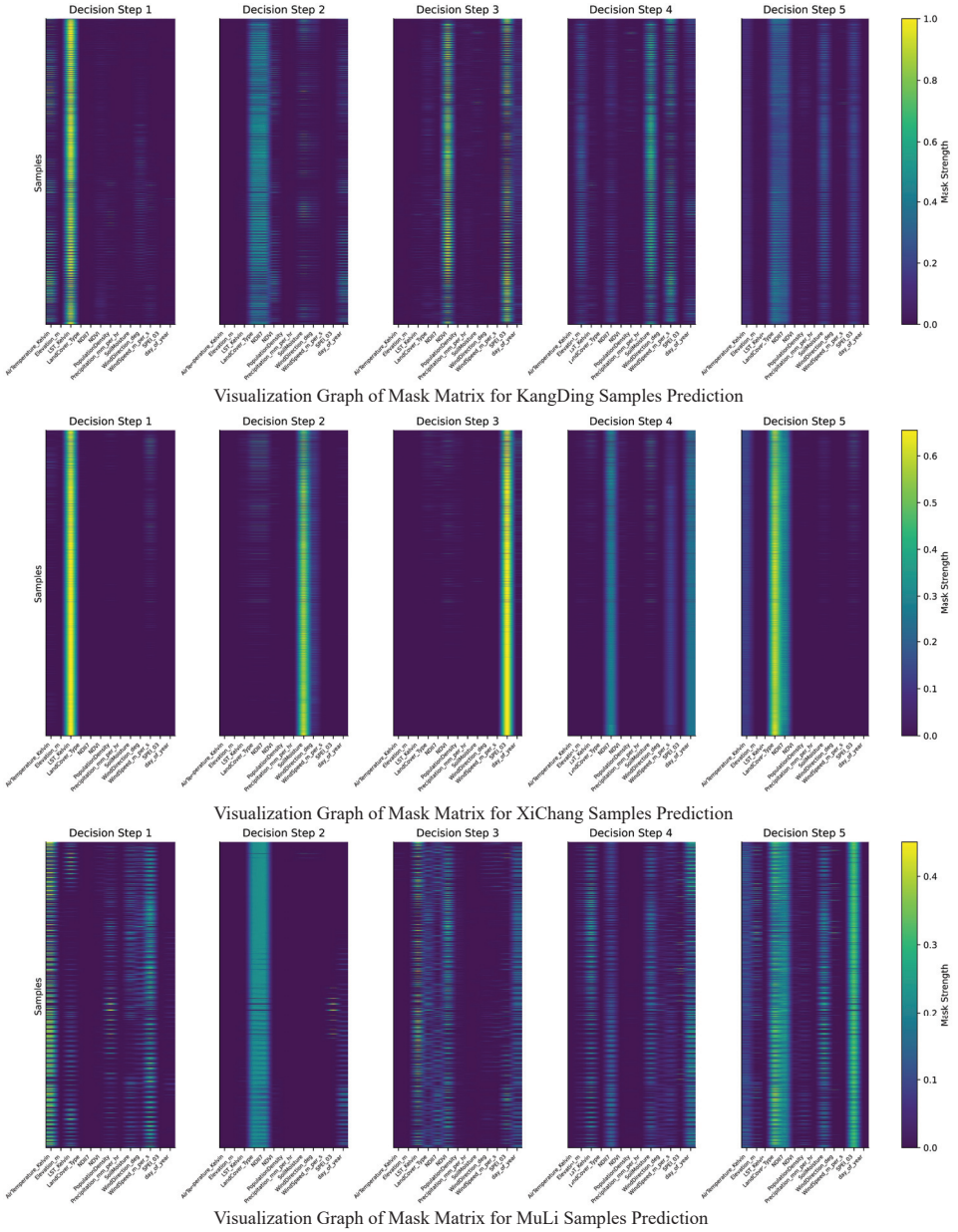


Figure 12. Mask visualizations for three wildfire incidents. Brighter bands indicate greater involvement of the corresponding feature in classification at each decision step. The x-axis labels for each step, from left to right, are: AirTemperature_Kelvin, Elevation_m, LST_Kelvin, LandCover_Type, NDVI, PopulationDensity, Precipitation_mm_per_hr, SoilMoisture, WindDirection_deg, WindSpeed_m_per_s, SPEI_03, and day_of_year.

https://doi.org/10.3390/fire9030107

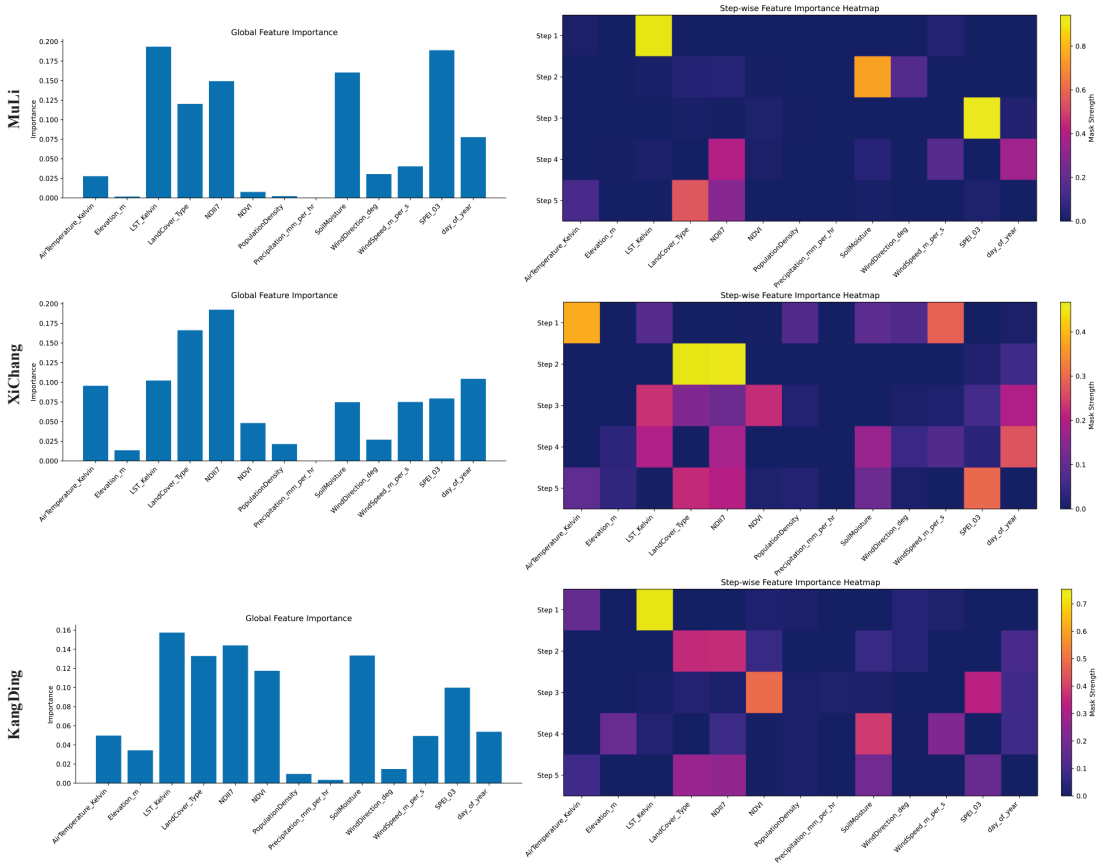


Figure 13. Global and step-wise feature importance for three wildfire incidents (MuLi, XiChang, and KangDing) derived from TabNet masks. For each incident, the left panel shows aggregated (global) feature importance, while the right panel presents the step-wise feature-mask strength across decision steps (Steps 1–5). Larger bars and warmer colors indicate higher contribution of the corresponding predictors (AirTemperature, Elevation, LST, LandCover, NDVI, PopulationDensity, Precipitation, SoilMoisture, WindDirection, WindSpeed, SPEI_03, and day_of_year) to the final classification.

Combining the spatial susceptibility maps (Figure 8) with the above interpretability results, Xichang—a typical mountainous area with pronounced dry–wet seasonality—illustrates that TabNet produces a more coherent and targeted depiction of localized risk in valleys, on slopes, and around settlements. In contrast, other models sometimes yield fragmented high-risk patches or miss key areas, which may reflect limitations in capturing multi-factor coupling under spatial heterogeneity.

Overall, the global and event-level masks indicate that TabNet can capture coupled environmental drivers while preserving geographically differentiated susceptibility signals, which is beneficial for operational risk zoning and for prioritizing monitoring and mitigation efforts.

4.3. Interpretable Methods and Their Applications

TabNet generates embedded feature masks through sequential sparse attention, while SHAP is a widely adopted post hoc interpretability method for tree-based models. Using the Kangding region as a case study (Figure 14), both approaches reveal similar overall patterns—thermal factors (such as LST and NDII7) and vegetation indices (like NDVI) show strong contributions, while precipitation has minimal impact. However, the methods diverge in their detailed rankings, such as the influence of population density. This discrepancy reflects the differing mechanisms of the models and their explanation frameworks: TabNet highlights dominant factors that are consistently selected across decision steps, while SHAP aggregates contributions from multiple tree paths, potentially emphasizing

https://doi.org/10.3390/fire9030107

features that frequently trigger localized splits. In practice, combining both methods can help identify the most stable environmental drivers, while still accounting for occasional human or extreme events.

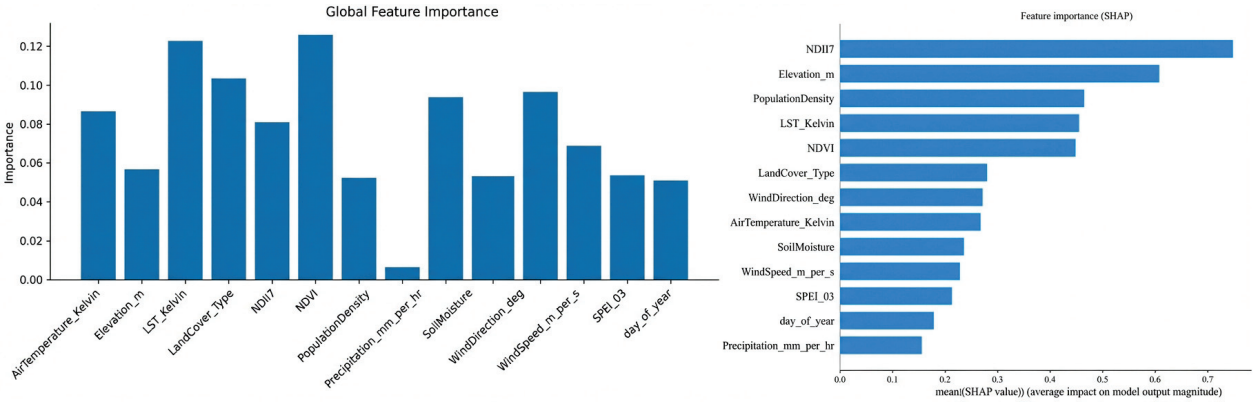


Figure 14. Comparison of interpretability methods in the Kangding incident: TabNet masks vs. XGBoost-SHAP. TabNet leverages sparse attention to generate embedded feature masks, highlighting dominant factors that are consistently selected during the decision-making process. SHAP, on the other hand, aggregates feature contributions across various decision paths in tree models. Both methods identify thermal factors (e.g., LST and NDII7) and vegetation indices (e.g., NDVI) as key drivers, while precipitation has a limited impact. Differences arise in the detailed ranking of features, such as population density, due to the models' differing mechanisms: TabNet emphasizes factors selected across decision steps, while SHAP focuses on aggregated contributions from multiple tree paths.

The interpretability framework proposed here sets itself apart from purely post hoc methods by leveraging TabNet's intrinsic sparse feature selection. By tracking the attention masks during inference, it provides a transparent decision trace and facilitates feature-space refinement (for example, by eliminating consistently low-importance variables). The reduced feature subset can then be transferred to traditional models, such as Random Forest or GBDT, improving efficiency and enabling cross-model comparison.

Practical Application Value

The interpretability of this model significantly enhances its practical application, particularly in fields like environmental monitoring, disaster management, and policy-making. By providing insights into the key factors driving environmental changes, this framework can help policymakers and environmental scientists prioritize areas that require attention, allocate resources more effectively, and predict potential future events. For instance, understanding the relative importance of factors like temperature, vegetation indices, and precipitation can guide more targeted interventions in regions prone to natural disasters, such as floods or droughts. Furthermore, the ability to refine feature space and transfer the reduced feature set to other models improves the model's adaptability and scalability, making it suitable for a wide range of applications, from climate modeling to urban planning.

5. Conclusions

This study developed an intrinsically interpretable wildfire-susceptibility assessment framework for forested landscapes in Southwest China by integrating multi-source environmental and anthropogenic predictors with the TabNet deep learning architecture. Motivated by the increasing frequency and severity of wildfires under climate change and intensifying human activities, the proposed approach addresses a key limitation of conventional machine-learning pipelines—namely, the trade-off between predictive accuracy and

https://doi.org/10.3390/fire9030107

interpretability—by providing built-in, step-wise explanations through TabNet's sparse feature-selection mechanism.

Using Google Earth Engine, we constructed a harmonized 1 km × 1 km dataset (2010–2020) that combines MODIS FireMask-derived labels with climatic (air temperature, precipitation, wind, soil moisture), vegetation (NDVI, NDII7, land cover), topographic (elevation), anthropogenic (population density), and seasonal (day-of-year) variables. A strict sampling strategy, including FireMask-based label filtering, a 5 km buffer to reduce spatial dependence, forest-land-cover masking, and data cleaning with Isolation Forest, was adopted to improve label reliability and reduce trivial separability. To mitigate class imbalance, class weighting and resampling strategies were incorporated during model training and evaluation.

The optimized TabNet model demonstrated strong wildfire-detection capabilities, achieving an AUC of 0.760 and a high recall (0.9577) on the independent test set, with an $F_{1.5}$ score of 0.7746, outperforming Random Forest, XGBoost, and LightGBM under the same preprocessing and validation protocol. Event-level case studies further indicated that high-susceptibility zones generally aligned with observed MODIS fire points, suggesting that the model can capture spatially coherent risk gradients in complex mountainous environments. Importantly, TabNet's mask matrices provided transparent global and event-specific interpretations, consistently highlighting thermal and vegetation-related controls (e.g., LST, land cover, NDVI) together with seasonality, while moisture and wind variables offered secondary but non-negligible contributions. The agreement between mask-based rankings and SHAP analyses further supports the reliability of the intrinsic explanations.

Overall, the proposed framework combines accuracy with operational transparency, offering a deployable pathway for wildfire early warning, risk zoning, and resource prioritization in forested regions. Future work should examine the sensitivity of susceptibility patterns to key assumptions (e.g., incorporating a buffer radius, land-cover masking, and sampling design).

Author Contributions: Conceptualization, C.M. and S.Y.; methodology, C.M. and J.C.; software, C.M. and D.Z.; validation, C.M., S.Y. and J.C.; formal analysis, C.M.; investigation, C.M., Q.L. and Q.Y.; resources, S.Y. and J.G.; data curation, C.M.,and J.C.; writing—original draft preparation, C.M. and C.Q.; writing—review and editing, C.M. and S.Y.; visualization, C.M.,X.W. and Q.L.; supervision, S.Y. and J.C.; project administration, S.Y.; funding acquisition, S.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Civil Aerospace Technology Advanced Research Project of China under Grant D040306.

Data Availability Statement: Data supporting the results of this study are available at https://github.com/machenyu2023/wildfire-dataset (accessed on 28 December 2025).

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

GEE	Google Earth Engine
NDVI	Normalized Difference Vegetation Index
NDII	Normalized Difference Infrared Index

https://doi.org/10.3390/fire9030107

References

1. Canadell, J.G.; Meyer, C.P.; Cook, G.D.; Dowdy, A.; Briggs, P.R.; Knauer, J.; Pepler, A.; Haverd, V. Multi-decadal increase of forest burned area in Australia is linked to climate change. *Nat. Commun.* **2021**, *12*, 6921. [CrossRef]
2. Kelly, L.T.; Giljohann, K.M.; Duane, A.; Aquilue, N.; Archibald, S.; Batllori, E.; Bennett, A.F.; Buckland, S.T.; Canelles, Q.; Clarke, M.F.; et al. Fire and biodiversity in the Anthropocene. *Science* **2020**, *370*, eabb0355. [CrossRef]
3. Byrne, B.; Liu, J.; Bowman, K.W.; Pascolini-Campbell, M.; Chatterjee, A.; Pandey, S.; Miyazaki, K.; van der Werf, G.R.; Wunch, D.; Wennberg, P.O.; et al. Carbon emissions from the 2023 Canadian wildfires. *Nature* **2024**, *633*, 835–839. [CrossRef]
4. Park, C.Y.; Takahashi, K.; Fujimori, S.; Jansakoo, T.; Burton, C.; Huang, H.; Kou-Giesbrecht, S.; Reyer, C.P.O.; Mengel, M.; Burke, E.; et al. Attributing human mortality from fire PM2.5 to climate change. *Nat. Clim. Change* **2024**, *14*, 1193–1200. [CrossRef]
5. Burke, M.; Childs, M.L.; de la Cuesta, B.; Qiu, M.; Li, J.; Gould, C.F.; Heft-Neal, S.; Wara, M. The contribution of wildfire to PM2.5 trends in the USA. *Nature* **2023**, *622*, 761–766. [CrossRef]
6. Kolden, C.A.; Abatzoglou, J.T.; Jones, M.W.; Jain, P. Wildfires in 2024. *Nat. Rev. Earth Environ.* **2025**, *6*, 367. [CrossRef]
7. Luz, A.E.O.; Negri, R.G.; Massi, K.G.; Colnago, M.; Silva, E.A.; Casaca, W. Mapping Fire Susceptibility in the Brazilian Amazon Forests Using Multitemporal Remote Sensing and Time-Varying Unsupervised Anomaly Detection. *Remote Sens.* **2022**, *14*, 2429. [CrossRef]
8. Üstek, İ.; Arana-Catania, M.; Farr, A.; Petrunin, I. Deep Autoencoders for Unsupervised Anomaly Detection in Wildfire Prediction. *Earth Space Sci.* **2024**, *11*, e2024EA003997. [CrossRef]
9. Barzani, A.R.; Pahlavani, P.; Ghorbanzadeh, O. Ensembling of decision trees, knn, and logistic regression with soft-voting method for wildfire susceptibility mapping. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2023**, *10*, 647–652. [CrossRef]
10. Jiang, W.; Qiao, Y.; Zheng, X.; Zhou, J.; Jiang, J.; Meng, Q.; Su, G.; Zhong, S.; Wang, F. Wildfire risk assessment using deep learning in Guangdong Province, China. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *128*, 103750. [CrossRef]
11. Wang, N.; Zhao, S.; Wang, S. A novel clustering-based resampling with cost-sensitive boosting method to model and map wildfire susceptibility. *Reliab. Eng. Syst. Saf.* **2024**, *242*, 109742. [CrossRef]
12. Rui, C.; He, B.; Li, Y.; Zhang, Y.; Liao, Z.; Fan, C.; Yin, J.; Zhang, H. Incorporating fire spread simulation and machine learning algorithms to estimate crown fire potential for pine forests in Sichuan, China. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *132*, 104080.
13. Sazib, N.; Bolten, J.D.; Mladenova, I.E. Leveraging NASA Soil Moisture Active Passive for Assessing Fire Susceptibility and Potential Impacts Over Australia and California. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 779–787. [CrossRef]
14. Ting, Y.; Li, J.; Ma, L.; Zhou, J.; Wang, R.; Eichhorn, M.P.; Zhang, H. Status, advancements and prospects of deep learning methods applied in forest studies. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *131*, 103938.
15. Nur, A.S.; Kim, Y.J.; Lee, J.; Lee, C.W. Spatial Prediction of Wildfire Susceptibility Using Hybrid Machine Learning Models Based on Support Vector Regression in Sydney, Australia. *Remote Sens.* **2023**, *15*, 760. [CrossRef]
16. Di Giuseppe, F.; McNorton, J.; Lombardi, A.; Wetterhall, F. Global data-driven prediction of fire activity. *Nat. Commun.* **2025**, *16*, 2918. [CrossRef]
17. Bui, D.T.; Le, K.T.T.; Nguyen, V.C.; Le, H.D.; Revhaug, I. Tropical Forest Fire Susceptibility Mapping at the Cat Ba National Park Area, Hai Phong City, Vietnam, Using GIS-Based Kernel Logistic Regression. *Remote Sens.* **2016**, *8*, 347.
18. Li, Y.; Xu, S.; Fan, Z.; Zhang, X.; Yang, X.; Wen, S.; Shi, Z. Risk Factors and Prediction of the Probability of Wildfire Occurrence in the China-Mongolia-Russia Cross-Border Area. *Remote Sens.* **2023**, *15*, 42. [CrossRef]
19. Sun, Y.; Zhang, F.; Lin, H.; Xu, S. A Forest Fire Susceptibility Modeling Approach Based on Light Gradient Boosting Machine Algorithm. *Remote Sens.* **2022**, *14*, 4362. [CrossRef]
20. Janizadeh, S.; Tran, T.T.K.; Bateni, S.M.; Jun, C.; Kim, D.; Trauernicht, C.; Heggy, E. Advancing the LightGBM approach with three novel nature-inspired optimizers for predicting wildfire susceptibility in Kaua`i and Moloka`i Islands, Hawaii. *Expert Syst. Appl.* **2024**, *258*, 124963. [CrossRef]
21. Xie, L.; Zhang, R.; Zhan, J.; Li, S.; Shama, A.; Zhan, R.; Wang, T.; Lv, J.; Bao, X.; Wu, R. Wildfire Risk Assessment in Liangshan Prefecture, China Based on An Integration Machine Learning Algorithm. *Remote Sens.* **2022**, *14*, 4592. [CrossRef]
22. Zhang, G.; Wang, M.; Liu, K. Deep neural networks for global wildfire susceptibility modelling. *Ecol. Indic.* **2021**, *127*, 107735. [CrossRef]
23. Kanwal, R.; Rafaqat, W.; Iqbal, M.; Song, W. Data-Driven Approaches for Wildfire Mapping and Prediction Assessment Using a Convolutional Neural Network (CNN). *Remote Sens.* **2023**, *15*, 5099. [CrossRef]
24. Rezaie, F.; Panahi, M.; Bateni, S.M.; Lee, S.; Jun, C.; Trauernicht, C.; Neale, C.M.U. Development of novel optimized deep learning algorithms for wildfire modeling: A case study of Maui, Hawai'i. *Eng. Appl. Artif. Intell.* **2023**, *125*, 106699. [CrossRef]
25. Iban, M.C.; Aksu, O. SHAP-Driven Explainable Artificial Intelligence Framework for Wildfire Susceptibility Mapping Using MODIS Active Fire Pixels: An In-Depth Interpretation of Contributing Factors in Izmir, Türkiye. *Remote Sens.* **2024**, *16*, 2842. [CrossRef]
26. Abdollahi, A.; Pradhan, B. Explainable artificial intelligence (XAI) for interpreting the contributing factors feed into the wildfire susceptibility prediction model. *Sci. Total Environ.* **2023**, *879*, 163004. [CrossRef]

https://doi.org/10.3390/fire9030107

27. Ramayanti, S.; Kim, B.; Park, S.; Lee, C.W. Wildfire susceptibility mapping by incorporating damage proxy maps, differenced normalized burn Ratio, and deep learning algorithms based on sentinel-1/2 data: A case study on Maui Island, Hawaii. *GIScience Remote Sens.* **2024**, *61*, 2353982. [CrossRef]
28. Arik, S. .; Pfister, T. TabNet: Attentive Interpretable Tabular Learning. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 6679–6687. [CrossRef]
29. Li, W.; Kang, J.; Wang, Y. Exploring the interactions and driving factors among typical ecological risks based on ecosystem services: A case study in the Sichuan-Yunnan ecological barrier area. *Ecol. Indic.* **2025**, *170*, 113000. [CrossRef]
30. Ying, L.; Cheng, H.; Shen, Z.; Guan, P.; Luo, C.; Peng, X. Relative humidity and agricultural activities dominate wildfire ignitions in Yunnan, Southwest China: Patterns, thresholds, and implications. *Agric. For. Meteorol.* **2021**, *307*, 108540. [CrossRef]
31. Du, J.; Wang, K.; Cui, B. Attribution of the Extreme Drought-Related Risk of Wildfires in Spring 2019 over Southwest China. *Bull. Am. Meteorol. Soc.* **2021**, *102*, S83–S90. [CrossRef]
32. Yin, J.; He, B.; Fan, C.; Chen, R. Fire has become a major disturbance agent in the forests of Southwest China. *Ecol. Indic.* **2024**, *160*. [CrossRef]
33. Xing, H.; Fang, K.; Yao, Q.; Zhou, F.; Ou, T.; Liu, J.; Zhou, S.; Jiang, S.; Chen, Y.; Bai, M.; et al. Impacts of changes in climate extremes on wildfire occurrences in China. *Ecol. Indic.* **2023**, *157*, 111288. [CrossRef]
34. Jiao, M.; Quan, X.; Yao, J.; Wang, W. How Does the Management Paradigm Contain Wildfire Over Southwest China? Evidence From Remote Sensing Observation. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 2504205. [CrossRef]
35. Liu, J.; Wang, Y.; Lu, Y.; Zhao, P.; Wang, S.; Sun, Y.; Luo, Y. Application of Remote Sensing and Explainable Artificial Intelligence (XAI) for Wildfire Occurrence Mapping in the Mountainous Region of Southwest China. *Remote Sens.* **2024**, *16*, 3602. [CrossRef]
36. Wilson, N.R.; Norman, L.M. Analysis of vegetation recovery surrounding a restored wetland using the normalized difference infrared index (NDII) and normalized difference vegetation index (NDVI). *Int. J. Remote Sens.* **2018**, *39*, 3243–3274. [CrossRef]
37. Giglio, L.; Loboda, T.; Roy, D.P.; Quayle, B.; Justice, C.O. An active-fire based burned area mapping algorithm for the MODIS sensor. *Remote Sens. Environ.* **2009**, *113*, 408–420. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

https://doi.org/10.3390/fire9030107

remote sensing

MDPI

Article

AMFA-DeepLab: An Improved Lightweight DeepLabV3+ Adaptive Multi-Statistic Fusion Attention Network for Sea Ice Segmentation in GaoFen-1 Images

Zengzhou Hao [1,2,3,*], Xin Li [1,4], Qiankun Zhu [1], Yunzhou Li [2,3], Zhihua Mao [1], Jianyu Chen [1] and Delu Pan [1,3]

[1] State Key Laboratory of Satellite Ocean Environment Dynamics, Second Institute of Oceanography, Ministry of Natural Resources, Hangzhou 310012, China; lixin23@zjou.edu.cn (X.L.); zhuqiankun@sio.org.cn (Q.Z.); mao@sio.org.cn (Z.M.); chenjianyu@sio.org.cn (J.C.); pandelu@sio.org.cn (D.P.)
[2] Institute of Oceanographic Instrumentation, Qilu University of Technology (Shandong Academy of Sciences), Qingdao 266061, China; lyz@qlu.edu.cn
[3] Academician Workstation of Shandong Province, Shandong Academy of Sciences, Jinan 250014, China
[4] School of Information Engineering, Zhejiang Ocean University, Zhoushan 316022, China
* Correspondence: hzyx80@sio.org.cn

Highlights

What are the main findings?

- By adopting MobileNetV2 as the backbone in DeepLabV3+, the model achieves extreme lightweight design with only 5.85 million parameters and the fast inference speed with 281.76 frames per second.
- By constructing an innovative adaptive multi-statistic fusion attention (AMFA) module, the model dynamically enhances image edge features while suppressing background interference for sea ice segmentation in GaoFen-1 images.

What are the implications of the main findings?

- A new AMFA-DeepLab is proposed by integrating the MobileNetV2 and AMFA modules in the DeepLabV3+ network model and is applied for sea ice segmentation in GaoFen-1 Images.
- The AMFA-DeepLab achieves high-precision segmentation of sea ice under complex marine conditions, with precision, recall, F1-score, and intersection over union reaching 94.82%, 97.03%, 95.91%, and 92.15%, respectively. It also maintains extremely low model complexity—only 10.7% that of the DeepLabV3+ model—while reducing training time by nearly 20%.

Abstract

For addressing difficult detail extraction and low operating efficiency in monitoring sea ice in a large area with wide-field-of-view images from the Chinese Gaofen-1 satellite, a lightweight, high-precision sea ice segmentation network adaptive multistatistic fusion attention (AMFA) module using DeepLabV3+ as the base architecture (AMFA-DeepLab) is proposed. First, the module replaces the backbone network with a lightweight MobileNetV2 to ensure feature extraction capability and greatly reduce model computational complexity using inverted residuals and depthwise separable convolution. Second, to solve the problems of fragmented ice texture blurring and speckle noise interference in optical images, an AMFA is designed and introduced into the decoder side. This module innovatively integrates the global median pooling branch and adapts the recalibrated feature weight through a dynamic channel mixing mechanism, effectively enhancing the model's capability of capturing fine sea ice edge features and its antinoise robustness in

complex backgrounds. Experimental results based on the dataset from Liaodong Bay in the Bohai Sea of China show that the intersection over union of AMFA-DeepLab reaches 92.15% and the F1-score reaches 95.91%, increases of 3.06%, and 1.68%, respectively, compared with those of the baseline model. In addition, only 5.85 million model parameters are needed, the training time is shortened to 4.42 h, and the inference speed is 281.76 frames per second. Visualized analysis and generalization test further demonstrates that this model can accurately eliminate clutter interference from coastal land and seawater and extract the fine filamentous structure of drift ice in the scene of complex melting ice. This research overcomes the precision bottleneck while achieving an ultimate lightweight model, providing efficient technical support for operational dynamic monitoring of sea ice disasters based on Chinese GaoFen-1 satellites.

Keywords: sea ice; lightweight network; DeepLab; attention module; GaoFen-1

1. Introduction

Changes in sea ice occupy an important position in the global climatic system. Sea ice not only influences global climatic patterns but also has far-reaching impacts on the ecosystem and shipping safety. On a global scale, China is one of the most impacted countries suffering from sea ice disasters. The occurrence of such disasters has close correlations with environmental conditions such as seawater temperature and sea area climate [1]. Affected by the dynamic features and potential risks of sea ice, some sailing activities face severe challenges that pose threats to sailing safety and ship operating efficiency [2]. Moreover, in the field of marine fisheries, floating sea ice on the sea may destroy aquaculture facilities in nearshore areas, thus resulting in the death of fish and shrimp owing to insufficient dissolved oxygen in the water [3]. Consequently, accurately and efficiently obtaining information on sea ice distribution has a key practical value for safeguarding marine shipping, fishery aquaculture, and the marine ecological environment. At present, the core technical means for sea ice monitoring include on-site icebreaker cruising, deployment of coastal observation stations, and satellite remote sensing. However, because sea ice is mostly distributed in high-latitude remote sea areas, the extreme weather conditions and complex marine conditions make traditional monitoring means costly and high-risk. With its rapid development, satellite remote sensing technology has become a mainstream means for sea ice monitoring in large areas owing to its wide coverage, excellent temporal resolution, and rapid data acquisition.

The development history of sea ice remote sensing image recognition technology entails three stages: traditional image processing, shallow machine learning, and deep learning. Each method has its advantages and disadvantages with respect to feature extraction and applicable scenes. In the early stages, traditional image processing methods primarily relied on the physical differences in grayscale, texture, and spectrum between sea ice, seawater, and land to achieve extraction through threshold segmentation or edge detection. For example, Liu et al. [4] and Zou et al. [5] completed the classification of sea ice in the Arctic and the Antarctic, respectively, by using differences in the backscattering coefficient and texture. Li et al. [6] enhanced the sea ice drifting monitoring capability of Synthetic Aperture Radar (SAR) images by texture feature screening. Such methods are characterized by directly viewing physical meaning and convenient computation, but their threshold setting relies highly on human experience, making it difficult to adapt to complex and varied marine conditions and lighting conditions. Afterward, shallow machine learning methods represented by support vector machines and random forests emerged gradually.

https://doi.org/10.3390/rs18050783

Zhou et al. [7] extracted sea ice parameters using FY-3 data by combining spectral and texture features. Li et al. [8] introduced a random forest algorithm to achieve automatic classification of sea ice under clouds with FY-4A data. Li et al. [9] achieved accurate extraction for sea ice in Liaodong Bay based on HY-1C/D data using multi-channel feature rules. However, such methods still need sophisticated artificial feature engineering, and the shallow models have difficulty in capturing high-dimensional semantic information. Consequently, their generalization capability is still limited in complex scenes with cloud and fog interference or vague sea ice–seawater boundaries.

In recent years, the breakthrough of deep learning technology has brought about a paradigm shift for remote sensing interpretation. Classic architectures such as fully convolutional networks [10] and U-Net [11] significantly improved sea ice recognition precision through their end-to-end feature-learning capability. At present, many scholars have proposed modified models for different data sources. Owing to their advantage of all-weather observation, SAR images are researched most widely. For example, Xu et al. [12] optimized the U-Net hybrid loss function for Sentinel-1 dual-polarized SAR data. Zhang et al. [13] constructed MSI-ResNet based on Gaofen-3 (GF-3) fully polarized SAR images. Dai et al. [14] proposed an improved IceDeepLab network for Arctic sea ice classification, demonstrating the effectiveness of the DeepLabV3+ architecture in processing large-scale SAR imagery. Sun et al. [15] validated the effectiveness of DeepLabV3+ in SAR images of the Greenland Sea. For high-spectral and multispectral images, Han et al. [16,17] explored deep spectral features of sea ice in Baffin Bay using three-dimensional convolutional and multiscale feature fusion networks. Furthermore, for modification of the universal model, Zhou et al. [18] and Zheng et al. [19] further improved the performance of YOLO and U-Net++ in sea ice monitoring in polar regions separately by introducing an attention module and modifying the convolution module.

Although many research results have been attained thus far, there remains an evident lag in the adaptation between data sources and algorithms in the field of sea ice monitoring based on deep learning. Current deep learning models mostly focus on SAR data or some optical satellite data. However, research on models customized to Chinese GF-1 satellite wide-field-of-view (WFV) sensors is relatively scarce. GF-1 WFV sensor observes 16 m-resolution broad-width images, which can concurrently achieve large coverage area and fine texture, making them very suitable for sea ice disaster monitoring in nearshore aquaculture areas.

Nevertheless, in contrast to SAR data's focus on backscattering, GF-1 data are characterized by the following: (1) sea ice in GF-1 high-resolution optical images being easily influenced by shielding from clouds and fog; (2) the presence of spectral confusion between thin ice and seawater; and (3) fractal features and fragmented ice details of sea ice edges being more complex owing to the increase in resolution. If existing models designed for SAR or medium- and low-resolution images are directly used, problems such as edge blur and missed detection of tiny floating ice blocks typically occur.

Among the numerous deep learning models, the DeepLabV3+ [20] model exhibits an extremely high potential for solving the above optical sea ice recognition problems owing to its unique structural advantage. Sea ice has a significant multiscale feature in remote sensing images, as they contain not only large-area fixed ice blocks but also fragmented tiny floating ice blocks. The Atrous Spatial Pyramid Pooling (ASPP) module introduced in the DeepLabV3+ model can effectively capture multiscale contextual information through atrous convolutions with different dilation rates. In addition, the encoder–decoder architecture of the model can effectively restore the spatial resolution, and it can better preserve the fine irregular boundaries of sea ice compared with the traditional fully convolutional network architecture. Drawing on its architectural advantages, DeepLabV3+ has exhibited

https://doi.org/10.3390/rs18050783

considerable potential for the fine-grained segmentation of diverse and complex terrestrial features. For example, Chen et al. [21] incorporated an axial attention mechanism into DeepLabV3+ to enhance the accuracy of oil spill detection in SAR imagery, while Feng et al. [22] and Yang et al. [23] further validated the efficacy of improved DeepLabV3+ architectures—through the introduction of multiscale attention and cross-scale feature interaction modules, respectively—in applications such as tailings pond identification and crop disease monitoring. Collectively, these studies underscore the strong capability of DeepLabV3+ as a baseline model for handling intricate textures and multiscale targets.

However, the DeepLabV3+ model still faces two key defects in business-oriented monitoring by directly adapting GF-1 WFV high-resolution images. The first is the contradiction between the computational redundancy and lightweight demand of the model. The original DeepLabV3+ model usually employs Xception [24] as its backbone network, which is characterized by numerous parameters and high computational complexity. Sea ice monitoring typically has an extremely high requirement for timeliness, the large GF-1 WFV width leads to too heavy a model, and the greatly increased data quantity makes is hard to achieve rapid processing of large-area images under limited computational power. The second is the missed detection of small targets caused by insufficient utilization of channel features. In GF-1 images, fragmented ice and thin ice exhibit little difference from the background seawater in terms of spectral features. The ASPP module of DeepLabV3+ mainly focuses on multiscale fusion on the spatial dimension and neglects the difference in feature response on the channel dimension. The lack of an effective channel attention module causes the model to fail to automatically suppress the response of the noise channel, making it difficult to focus on the key channel expressing the fragmented ice texture, finally resulting in confusing the sea ice with the complex background.

Therefore, to meet the dual requirements of sea ice detail extraction precision and real-time processing timeliness for GF-1 WFV data, a lightweight, high-precision sea ice segmentation network AMFA module using DeepLabV3+ as the base architecture (AMFA-DeepLab) is proposed. The two main modifications are the following: (1) For the computational efficiency, a lightweight MobileNetV2 network is used to replace the original computation-intensive Xception backbone network. With its unique inverted residuals and depthwise separable convolution, the network greatly reduces the number of model parameters and GPU memory usage while preserving the efficient extraction capability for multiscale features of sea ice. (2) To address the difficult problem of fine feature recognition, an AMFA module is designed and introduced in the decoder feature fusion stage. In contrast to a routine single pooling strategy, the module innovatively integrates average, maximum, and global median pooling and generates adaptive channel weights to recalibrate the features through a dynamic mixing mechanism. In particular, the introduced median pooling significantly enhances the robustness of the model against speckle noise in remote sensing images so that the model can better capture fine sea ice textures and effectively suppress the interference from complex backgrounds. Through the above modifications, AMFA-DeepLab achieves the best balance between high precision and light weight, providing effective technical support for business-oriented dynamic monitoring of large-area sea ice based on Chinese high-resolution satellites.

2. Materials and Methods

2.1. Base Architecture of DeepLabV3+ and Its Limitations

DeepLabV3+ is a highly representative encoder–decoder framework used in image semantic segmentation. This model is based on DeepLabV3 and introduces an ASPP module. It can efficiently capture multiscale contextual information through atrous convolutions with different dilation rates, effectively solving the difficult problem of distinguishing sea

https://doi.org/10.3390/rs18050783

ice blocks of great differences in size. In addition, the decoder module can better restore sea ice edge details by fusing bottom-level spatial features and high-level semantic features. However, the original DeepLabV3+ usually uses Xception as its backbone network, and it has evident limitations in large-area sea ice monitoring based on GF-1 WFV data. Because the Xception network is characterized by numerous parameters and high computational complexity, meeting the harsh requirement for real-time capability in emergency monitoring for sea ice disasters is challenging. Moreover, the Xception network lacks self-adaption in the weight distribution for feature channels; consequently, segmentation precision easily decreases because of noise accumulation in complex marine conditions with cloud and fog interference.

2.2. The Overall Framework of the AMFA-DeepLab Network

The modification of the DeepLabV3+ network in this study is primarily embodied in the following two aspects: (1) The backbone network Xception module is replaced with a lightweight MobileNetV2 module. With its unique inverted residuals introduced into the linear bottleneck layer, Xception significantly reduces the number of parameters and computational complexity of the model while ensuring sea ice feature extraction precision. (2) An AMFA module is introduced after the fusion of high- and lower level features of the decoder. This module introduces the global median pooling branch against the speckle noise interference common in sea ice remote sensing images on the basis of traditional average pooling and maximum pooling and adaptively adjusts the weights of all statistics through dynamic channel mixing to conduct secondary refining of the fused feature images. Experimental results show that the AMFA module can effectively suppress incorrect activation of non-ice areas, enhancing the capability of the model for capturing fragmented ice and edge details. The modified model provides more accurate solutions for large-area sea ice monitoring while preserving efficient reasoning. The overall architecture of the AMFA-DeepLab network model proposed here is shown in Figure 1.

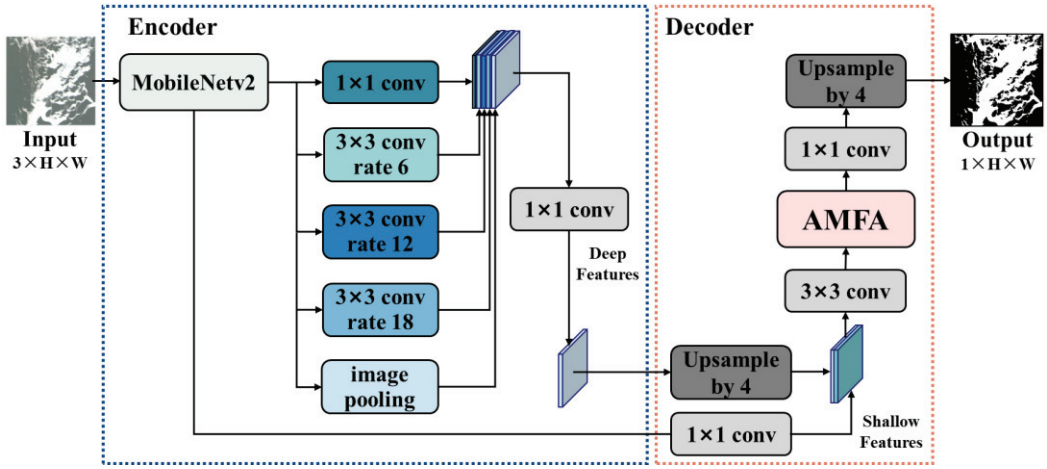


Figure 1. Framework of the AMFA-DeepLab network model.

2.2.1. MobileNetV2 Module

To balance the precision and computational efficiency of the model in sea ice feature extraction, we selected MobileNetV2 module as the backbone network. As a lightweight convolutional neural network specially designed for mobile terminals, the core innovation of MobileNetV2 lies in introducing inverted residuals into the linear bottleneck layer. As shown in Figure 2, this architecture exhibits two forms by convolution stride, each of which contains three key feature transformation stages. The first is the dimension expansion stage, in which the feature channels are expanded using 1×1 convolution and

https://doi.org/10.3390/rs18050783

the low-dimensional features are mapped to a high-dimensional space with the ReLU6 activation function. The second is the depthwise separable convolution stage, in which spatial feature extraction is conducted channel by channel through 3 × 3 convolution kernels to greatly reduce the number of parameters and computational cost. The third is the dimension reduction stage, in which the features are projected to a low-dimensional space through 1 × 1 convolution. This architecture removes nonlinear activation in the end dimension reduction layer and constitutes a linear bottleneck layer with linear activation to avoid information loss. Furthermore, when the stride is 1 (the upper half in Figure 2), the architecture introduces a shortcut connection similar to that of ResNet to conduct addition for input and output element by element, forming inverted residuals. When the stride is 2 (the lower half in Figure 2), downsampling is conducted, and the shortcut connection is not used. Because fine features such as thin ice texture and fragmented ice edges are relatively fragile in high-dimensional manifolds, the introduced linear bottleneck layer and inverted residuals can effectively preserve the edge details and texture structures of sea ice to a maximum extent, enhancing the analytical ability of the model for complex ice conditions.

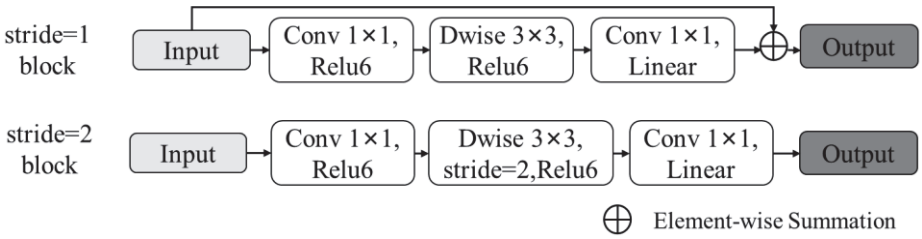


Figure 2. Inverted residual structure of the MobileNetV2 module.

2.2.2. AMFA Module

A channel attention module is generally used to improve model sensitivity to key features by explicitly modeling the interdependence between channels. The classic Squeeze-and-Excitation module [25] captures background statistical information using only global average pooling, whereas the convolutional block attention module [26] further introduces global maximum pooling to enhance the extraction of significant features. Nevertheless, in processing high-resolution sea ice remote sensing images, the above two methods have two shortcomings. The first is that high-frequency noise of sea surface wave light easily interferes with maximum pooling, resulting in incorrect activation of non-ice areas. The second is that the existing feature fusion mostly uses simple addition or splicing, lacking the ability to dynamically adjust the weights of different statistics based on image contents.

Consequently, we propose an AMFA module to reduce the impact of these two shortcomings. The AMFA module follows a three-stage architecture of complementary statistics extraction, nonlinear feature enhancement, and adaptive dynamic fusion. Specifically, this module is embedded into a node following the deep and shallow feature fusion stage of the DeepLabV3+ decoder. The fused features will contain the high-level semantic information and the low-level spatial details. Applying an attention module here can refine the pixel-level boundaries between sea ice and seawater while most effectively suppressing shallow-feature background noise. The structure of the AMFA module is shown in Figure 3.

When the input feature map after decoder fusion is $X \in \mathrm{R}^{B\times C\times H\times W}$, the AMFA module first compresses the spatial dimension through three parallel global pooling operations, generating three channel descriptors, which separately capture the distribution characteristics of the feature map on different statistical dimensions. The global average pooling (F_{avg}), extracting the smooth statistical information of the background, reflects the overall activation level of the feature channel. The global maximum pooling ($F_{\max}$), extracting the most significant texture and edge features, focuses on the key target area. The global

https://doi.org/10.3390/rs18050783

median pooling (F_{med}) aims at the distribution of salt-and-pepper noise specific to sea ice images. Unlike global average and maximum pooling, the global median pooling has natural statistical robustness against isolated high-frequency noises while preserving the representative central trend of feature distribution. In this study, we innovatively introduce median pooling to effectively filter isolated high-frequency noise, thereby compensating for the shortcoming that maximum pooling is sensitive to noise. Afterward, the above three descriptors perform nonlinear mapping separately through a Multi-Layer Perceptron (MLP), with weights shared to yield transformed high-dimensional feature vectors, $V_{avg}, V_{\max}, V_{med} \in R^{B\times C\times 1\times 1}$. The weight sharing ensures that the three statistical channels transform in the same mapping space, ensuring fairness in subsequent fusion and avoiding the introduction of too many additional parameters.

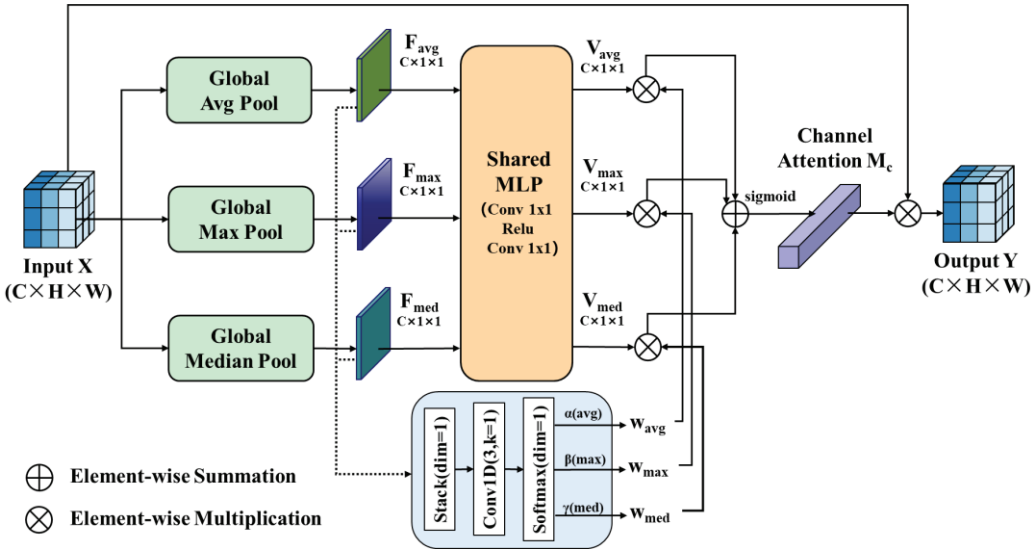


Figure 3. Structure of the AMFA module.

To achieve adaptive selection of features, we designed a dynamic channel mixer. In contrast to the static summation strategy of a traditional method, the strategy of this mixer is to dynamically generate normalized weights by learning the interrelationships of three statistics. First, the mixer stacks the original three pooling descriptors on dimension and captures the dependence between different statistics through a one-dimensional ($1D$) convolutional layer. Second, it uses the $Softmax$ function to perform the normalization on the statistic dimension, generating adaptive weight vectors:

$$W = Conv1D\big(Stack\big(F_{avg}, F_{\max}, F_{med}\big)\big), \tag{1}$$

$$[\alpha, \beta, \gamma] = Softmax(W), \tag{2}$$

where $\alpha, \beta, \gamma \in R^{B\times C}$ represents the importance weights of average, maximum, and median features on the current channel, respectively, and satisfies $\alpha + \beta + \gamma = 1$. Normalization constraints ensure that the fused features maintain a consistent numerical scale, preventing uncontrolled increase in feature amplitude. Finally, it uses the generated dynamic weights to perform weighted fusion of the MLP-transformed features, generating the final channel attention map, M_c, through the Sigmoid activation function, achieving recalibration of the original feature "X":

$$M_c = \sigma\big(\alpha \cdot V_{avg} + \beta \cdot V_{\max} + \gamma \cdot V_{med}\big), \tag{3}$$

$$Y = M_c \otimes X, \tag{4}$$

https://doi.org/10.3390/rs18050783

where σ is the Sigmoid function and $\otimes$ denotes multiplication, element by element.

In summary, the three mechanisms in the AMFA module contribute to the improvement of model performance at different levels. Firstly, the global median pooling enhances noise robustness by filtering out isolated high-frequency noises, and it can solve the problem of false activation in global max pooling within sea ice segments. Secondly, the shared MLP achieves nonlinear feature enhancement across statistical measures with minimal parameter overhead. The dynamic channel mixer enables the network to adaptively emphasize the most informative statistics on each channel based on image content and overcomes the limitations of traditional fixed fusion strategies. Through the above processing, the AMFA module can preserve the key texture features of sea ice to a maximum extent while suppressing background noise, thus significantly improving the robustness of the model in complex marine conditions.

2.2.3. Loss Function

In sea ice semantic segmentation tasks, there is typically a dramatic imbalance of pixel quantity between sea ice (foreground) and seawater (background); consequently, the traditional single-pixel-level loss function easily leads to overfitting of the model to the background. To solve this problem, a hybrid loss function constructed from a weighted combination of binary cross entropy (BCE) and the Dice coefficient [27] was used in this study. The total loss function L_{total} is defined as follows:

$$L_{total} = L_{BCE} + \lambda L_{Dice}, \tag{5}$$

where λ is a hyperparameter used to balance the weights of two loss terms, and it is set to 1.0 in this study. BCE loss (L_{BCE}) is used to measure pixel-level classification accuracy, ensuring the convergence stability of the model. Owing to its well-behaved log-likelihood gradient, BCE loss provides stable and reliable gradients throughout the entire training process, which is particularly important for compensating the gradient instability that Dice loss may exhibit during early training stages. It can be calculated as follows:

$$L_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \cdot \log(\hat{y}_i) + (1 - y_i) \cdot \log(1 - \hat{y}_i)], \tag{6}$$

where N denotes the total number of pixels in the batch, $y_i \in \{0, 1\}$ denotes the true label of the i-th pixel (where 0 represents the background and 1 represents the sea ice), and $\hat{y}_i \in [0, 1]$ denotes the value of the probability that the model predicts this pixel as sea ice.

Dice loss (L_{Dice}) originates from the field of medical image segmentation. It is used to directly optimize the overlap degree of sets and has high robustness for foreground–background class imbalance. Since Dice loss evaluates the overlap ratio between predicted and ground truth regions rather than relying on absolute pixel counts, it is inherently invariant to class proportion changes and can automatically strengthen the learning signal for the minority class regardless of the ice-water ratio in different images. It can be calculated as follows:

$$L_{Dice} = 1 - \frac{2\sum_{i=1}^{N} y_i \hat{y}_i + \epsilon}{\sum_{i=1}^{N} y_i + \sum_{i=1}^{N} \hat{y}_i + \epsilon}, \tag{7}$$

where the numerator denotes the intersection of the predicted map and the true value map, the denominator denotes the union of the above two, and ϵ is a smooth term used to prevent the denominator from being zero (and is generally taken as 1×10^{-5}). Through the combined optimization of the above two losses, the model can focus on pixel-level classification correctness and ensure the overall geometric integrity of the sea ice area. Therefore, BCE loss ensures convergence stability, while Dice loss provides ratio-invariant overlap optimization. The two losses form an effective complementary pair under equal

https://doi.org/10.3390/rs18050783

weighting (λ = 1.0), providing sufficiently robust optimization for the current sea ice segmentation without the need for additional dynamic weighting mechanisms.

2.3. Network Training

2.3.1. Dataset

The experimental data selected for this study were from the WFV sensors carried by the Chinese GF-1 satellite, and the study area was the Liaodong Bay in the Bohai Sea, China, which poses considerable risk of sea ice disasters and has a long icing period in winter and a wide distribution of drift ice, making it an ideal scene for verifying the robustness of the sea ice recognition algorithm (see Figure 4). In order to ensure the spatial-temporal diversity and representativeness of the dataset, the large-width GF-1 WFV L1A original images were obtained in January and February, from 2014 to 2016. These selected images cover a variety of typical sea ice conditions, from thick ice in the icing period to broken ice floes in the melting period. Because the GF-1 WFV sensors have a 16 m spatial resolution and 800 km coverage width, they can effectively capture the extremely rich and heterogeneous sea ice texture features, which is enough to meet the training needs of deep learning. To ensure the repeatability of study, the standardized preprocessing process is implemented. First, in the data preprocessing stage, red, green, and blue bands of the WFV images were selected to synthesize true color images, so as to preserve the most direct viewing visual texture features of sea ice. Second, to construct high-quality training labels, the sea ice area was defined as the foreground (with a pixel value of 1), and the seawater, land, and cloud layer were collectively defined as the background (with a pixel value of 0); the annotation strategy combined semiautomatic threshold extraction, supplemented by morphological operation for initial segmentation with artificial expert check, thus building pixel-level binary semantic segmentation true values. Third, the original remote sensing images were processed using the overlapping sliding window cropping strategy. The crop size and sliding step were set to 512 × 512 and 256 pixels, respectively. Finally, in this study, 3066 pairs of image-label samples in total were constructed, with the principle of random sampling being followed. Based on a ratio of 7:2:1, 3066 pairs were divided into three sets (i.e., a training set containing 2146 pairs, a validation set containing 613 pairs, and a training set containing 307 pairs), providing a data basis for subsequent model training and performance evaluation.

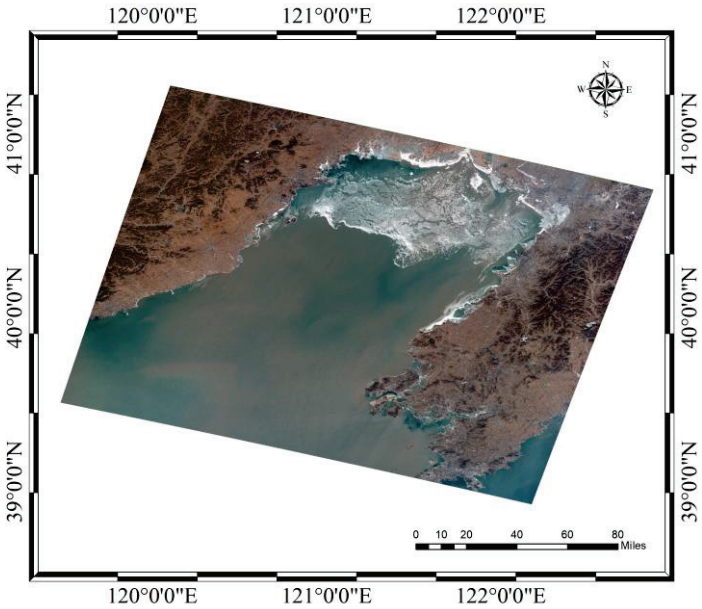


Figure 4. The typical GF-1 WFV raw image covering the Liaodong Bay on 10 January 2016.

https://doi.org/10.3390/rs18050783

2.3.2. Experimental Setting

All models in this study were constructed based on the PyTorch v2.5.1 deep learning framework and were subjected to accelerated training on an NVIDIA RTX 4090 GPU (with 24 GB GPU memory). An Adam optimizer was selected for parameter update, with the initial learning rate being set to 1×10^{-4} and the batch size being set to 8. To avoid getting stuck in local optima and improve model convergence, the cosine annealing learning rate schedule [28] was used to dynamically adjust the learning rate. To suppress the overfitting phenomenon and improve the robustness of the model generalization, online data augmentation, such as horizontal flip, vertical flip, and random scaling and photometric transformation, were introduced during training. Under the above configuration, the models were trained for 200 epochs in total and all of them tended to become stable and reach convergence states.

2.4. Evaluation Metrics

To comprehensively evaluate the overall performance of the modified model in sea ice semantic segmentation tasks, six evaluation metrics were selected, covering the segmentation precision and computational efficiency of the model. First, based on a pixel-level confusion matrix (including true positive (TP), false positive (FP), false negative (FN), precision (P), recall (R), F1-score ($F1$), and intersection over union (IoU)), Ref. [29] were selected to quantify the accuracy of sea ice extraction. Precision reflects the reliability of the prediction results, recall reflects the sea ice detection ratio of the model, the F1-score is the integrated harmonic mean of the above two, and IoU is used to measure the degree of overlap of the predicted and true areas; these are the most critical metrics for semantic segmentation. These metrics are calculated as follows:

$$P = \frac{TP}{TP + FP}, \tag{8}$$

$$R = \frac{TP}{TP + FN}, \tag{9}$$

$$F1 = 2 \times \frac{P \times R}{P + R}, \tag{10}$$

$$IoU = \frac{TP}{TP + FP + FN}, \tag{11}$$

where TP denotes the quantity of pixels that are correctly predicted as sea ice, FP denotes the quantity of background pixels that are incorrectly identified as sea ice, and FN denotes the quantity of sea ice pixels undetected. Then, owing to special requirements for real-time capability and hardware sources in sea ice disaster monitoring, model parameters training time and inference speed were introduced as efficiency evaluation metrics. The model parameters count the total quantity of all trainable weights (in millions), directly reflecting the spatial complexity of the model. The training time records the total time consumed by the model to complete the whole process of convergence on the training set (in hours) and is used to evaluate the computational cost and deployment potential of the algorithm. Finally, the inference speed is measured in frames per second (FPS), which calculates the average number of images processed by the model per second, serving as a key indicator to evaluate the real-time performance of the algorithm.

https://doi.org/10.3390/rs18050783

3. Results

3.1. The Results of the Ablation Experiments

To validate the effectiveness of the lightweight backbone network, MobileNetV2, and the AMFA module proposed in this study in sea ice segmentation tasks, step-by-step ablation experiments were designed and conducted on the sea ice dataset with the original DeepLabV3+ backbone (Xception backbone) as the baseline. The main comparison metrics for the experiment included P, R, F1, and IoU. The specific results are listed in Table 1.

Table 1. The results of the evaluation metrics for the ablation experiment.

Exp.	Module			P (%)	R (%)	F1 (%)	IoU (%)
	Xception	MobileNetV2	AMFA				
1	√	-	-	92.72	95.79	94.23	89.09
2	-	√	-	95.22	96.42	95.81	91.96
3	-	√	√	94.82	97.03	94.86	90.22

Where "√" means that the module is adopted, "-" means that the module is not adopted.

It can be seen from Table 1 that, in Experiment 1, Xception with a relatively large number of parameters is used as the backbone network, and IoU is 89.09%. In Experiment 2, the backbone network is replaced with a lightweight MobileNetV2 network. To ensure that the high-level semantic features retain sufficient spatial resolution after backbone replacement, the output stride was maintained at 16, consistent with the original DeepLabV3+ configuration. Specifically, for the convolutional blocks beyond the fourth downsampling stage, the original stride of 2 was replaced with a stride of 1, and atrous (dilated) convolutions with a dilation rate of 2 were applied to compensate for the reduced receptive field, thereby preserving the spatial detail of high-level feature maps. The low-level features were extracted from the output of the fourth convolutional block (at stride 4 relative to the input), which are subsequently fed into the decoder module for multiscale feature fusion. Experimental results show that the segmentation performance of the model does not decrease with the significant decrease in the number of parameters. Instead, IoU and F1 increased to 91.96% and 95.81%, respectively. This counterintuitive increase indicates that, in cases in which the sea ice samples are relatively limited and the texture features are specific, the over-deep and over-wide Xception network may suffer from overfitting or feature redundancy, and the MobileNetV2 network with its more compact architecture can more efficiently extract effective features through its linear bottleneck layer architecture, providing a good lightweight basis for subsequent module embedding. In Experiment 3, the AMFA module constructed in this study is introduced on the basis of a lightweight DeepLabV3+ model. Compared with Experiment 2, although the precision fluctuates slightly in Experiment 3, R further increases from 96.42% to 97.03%, indicating an enhanced sensitivity of the model to missed targets. The most critical integrated metric, IoU, finally reaches 92.15% in Experiment 3 (with an increase of 0.19% compared with Experiment 2 and of 3.06%, compared with the baseline), and F1 reaches 95.91%. This fully demonstrates that the AMFA module does not involve simple parameter stacking but, through adaptive weighting of channel attention, effectively compensates for the potential shortages of the lightweight network in capturing such fine textures as thin ice and cracks.

To visually explore the influences of the module on model attention, Figure 5 visualizes the feature response heatmaps of the model in different ablation experiment stages. Each row represents a typical scene (from the top down: dense fragmented ice blocks, nearshore land, and complex drift ice). Each column represents a model in a different stage (from left to right: original image, Xception baseline, lightweight MobileNetV2, and

https://doi.org/10.3390/rs18050783

MobileNetV2+AMFA). The redder the color is, the higher the confidence of the model for sea ice class will be; the bluer the color is, the stronger the background suppression will be.

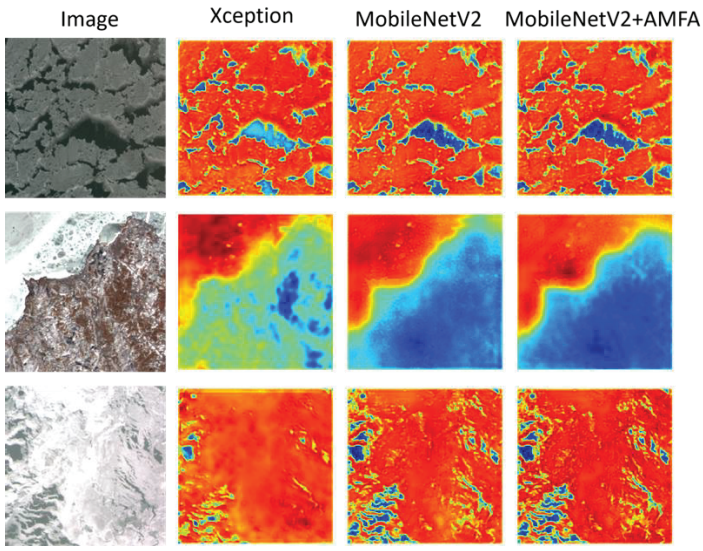


Figure 5. Visual comparison of feature response heatmaps in ablation experiment stages.

In the first row (dense fragmented ice blocks) and the third row (complex drift ice), heatmaps based on MobileNetV2+AMFA exhibit extremely high edge sharpness in the relatively scattered and vague activated areas based on the baseline model. The sea ice areas exhibit highlighted and uniform red activation, whereas the ice cracks and seawater gaps are clearly depicted as blue tones. Such high contrast of distinct red and blue indicates that the AMFA module successfully leads the model to concentrate computational resources onto the core sea ice textures, dramatically improving the significance and confidence of features. In the nearshore land scene in the second row, it can be discerned that, in the heatmaps based on Xception and MobileNetV2 networks, the land area (on the lower-left corner) exhibits large areas of yellow and light green clutter, indicating that confusion occurs in the model's feature extraction, with land and sea ice areas difficult to distinguish from each other. Comparatively, the heatmap based on the model with the AMFA module introduced exhibits pure dark blue in this area. This powerfully demonstrates that the median pooling branch in the AMFA module effectively filters off non-ice interference, including land textures, achieving perfect background suppression.

In summary, the visualized results corroborate the quantitative metrics for the ablation experiment from the perspective of interpretability: the AMFA module not only enhances the sensitivity of the model to sea ice but also greatly improves the recognition robustness of the model in complex backgrounds.

3.2. The Results of the Comparison Experiments

To compare the performance of different models, the following image segmentation networks were selected: SegNet, U-Net++ [30], U-NetV2 [31], MK-UNet [32], BiSeNetV2 [33], DeepLabV3+, and the AMFA-DeepLab proposed in this study. The evaluation metrics included P, R, F1, IoU, parameters, training time, and FPS.

To further validate the optimization efficiency and training stability of the modified model, Figure 6 shows the training loss convergence curves of the AMFA-DeepLab proposed in this study and of SegNet, U-Net++, U-NetV2, MK-UNet, BiSeNetV2, and the original DeepLabV3+ in 200 epochs. As can be clearly seen from Figure 6, the AMFA-DeepLab (red curve) exhibits steeper gradient during the initial stage of training, compared with modified networks such as U-Net++ (gray curve), U-NetV2 (cyan curve), and BiSeNetV2

https://doi.org/10.3390/rs18050783

(green curve). Although MK-UNet (purple curve) and U-NetV2 (cyan curve) perform better in the late stage than SegNet and U-Net++, the AMFA-DeepLab still maintains an evident advantage and finally stabilizes in the lowest loss value interval. The proposed AMFA-DeepLab is significantly better than the original DeepLabV3+ and other U-Net variants, indicating that the AMFA module more effectively captures the fine texture features of sea ice through multistatistic fusion.

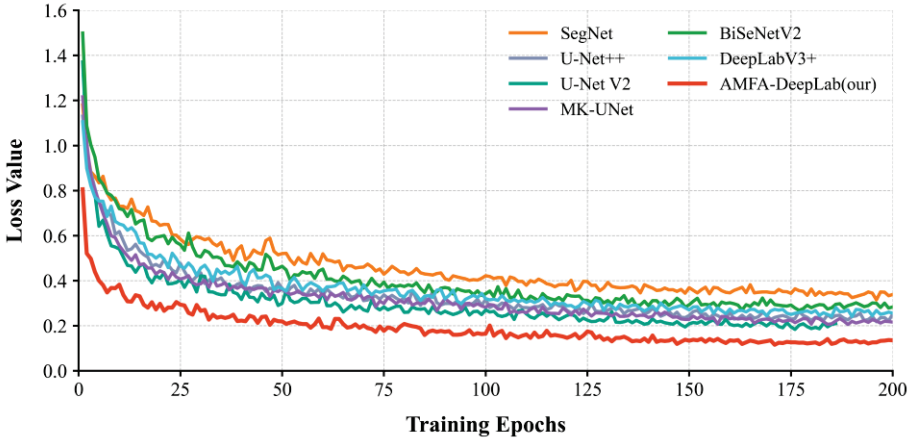


Figure 6. Comparison of loss function convergence curves during training among different models.

Table 2 summarizes the detailed evaluation metrics for different models on the test set. The experimental results show that the AMFA-DeepLab proposed in this study gives the best results with respect to all key precision metrics. Specifically, the IoU and F1 of AMFA-DeepLab reach 92.15%, and 95.91%, respectively, and the P and R also maintain the highest level. Compared with U-Net++ (IoU 89.95%), U-NetV2 (IoU 90.07%), and MK-UNet (IoU 89.38%), the modified model exhibits an evident performance advantage, with IoU increasing by 2.20%, 2.08%, and 2.77%, respectively. In addition, compared with the baseline model, DeepLabV3+, the AMFA-DeepLab exhibits a 3.06% increase in precision. This result powerfully demonstrates that the AMFA module, through multistatistic fusion, can more accurately capture the fine texture and edge features of sea ice than the traditional network, thus dramatically reducing both the missed detection rate and false detection rate. In terms of computational efficiency, the AMFA-DeepLab exhibits an excellent lightweight advantage. The number of parameters of the AMFA-DeepLab is only 5.85 million, which is far lower than that of DeepLabV3+ (54.71 million), only ~60% of that of U-Net++ (9.95 million) and equivalent to that of BiSeNetV2 (5.19 million). Moreover, the training time of the model is only 4.42 h, which is nearly 47% shorter than that of U-Net++ (8.34 h). It is worth noting that in terms of reasoning speed, based on 512 × 512 input, BiSeNetV2 achieves a higher frame rate (430.86 FPS), but its segmentation accuracy IOU is only 87.80%. In contrast, AMFA-DeepLab achieves the highest accuracy of all comparison models while maintaining the super real-time speed of 281.76 FPS. This demonstrates that the modified model significantly increases data processing speed throughout while greatly reducing the demand for computational power, making it even more suitable for the demand for timeliness in business-oriented monitoring of large-area sea ice.

https://doi.org/10.3390/rs18050783

Table 2. The results of experimental evaluation metrics for different models.

Model	P (%)	R (%)	F1 (%)	IoU (%)	Parameters (Millions)	Training Time (h)	Inference Speed (FPS)
SegNet	91.72	93.07	92.39	85.85	29.44	6.78	121.23
MK-UNet	92.81	96.02	94.39	89.38	5.24	8.69	142.42
BiSeNetv2	92.11	94.94	93.50	87.80	5.19	3.75	430.86
U-Net++	93.76	95.67	94.71	89.95	9.95	8.34	107.91
U-Netv2	93.54	96.04	94.77	90.07	25.15	4.44	95.86
DeepLabV3+	92.72	95.79	94.23	89.09	54.71	5.48	166.83
AMFA-Deepleb	**94.82**	**97.03**	**95.91**	**92.15**	5.85	4.42	281.76

where the bold formatting indicates the optimal value in each column.

Figure 7 shows the distribution relationships of parameters (x axis) and IoU precision (y axis) of all these models. The x value of each bubble in Figure 7 corresponds to the parameter scale of the corresponding model. As shown in the Figure 7, the AMFA-DeepLab (red bubble) is independently located in the uppermost left area of the coordinate system. This location indicates that this model simultaneously has the lowest model complexity and the highest segmentation precision. Compared with other models in Figure 7, especially the lightweight BiSeNetV2 (green bubble) and MK-UNet (purple bubble), which are also located in the left low parameter area, AMFA-DeepLab not only achieves significant compression of parameters in the x axis but even overcomes the precision bottleneck of existing models in the y axis. In addition, compared with the original DeepLabV3+, it achieves higher accuracy with a significant reduction in the number of parameters. This powerfully demonstrates that combining the linear bottleneck layer of MobileNetV2 with an AMFA module can effectively resolve the inherent contradiction in the traditional network that lightweight design decreases precision, providing the best solution for real-time accurate monitoring of sea ice based on high-resolution remote sensing.

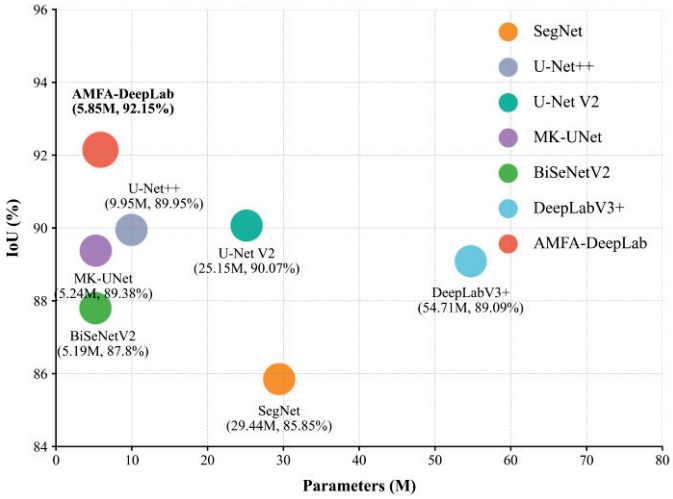


Figure 7. Comparison of IoU precision and efficiency of different segmentation networks.

To evaluate the detail capture capability of the model under different marine conditions, Figure 8 shows a visualized comparative analysis of four representative typical scenes selected from the test set. Each block image has a size of 512 × 512 pixels with 16 m spatial resolution. The original image, label, and prediction results based on various models are shown in turn from left to right.

https://doi.org/10.3390/rs18050783

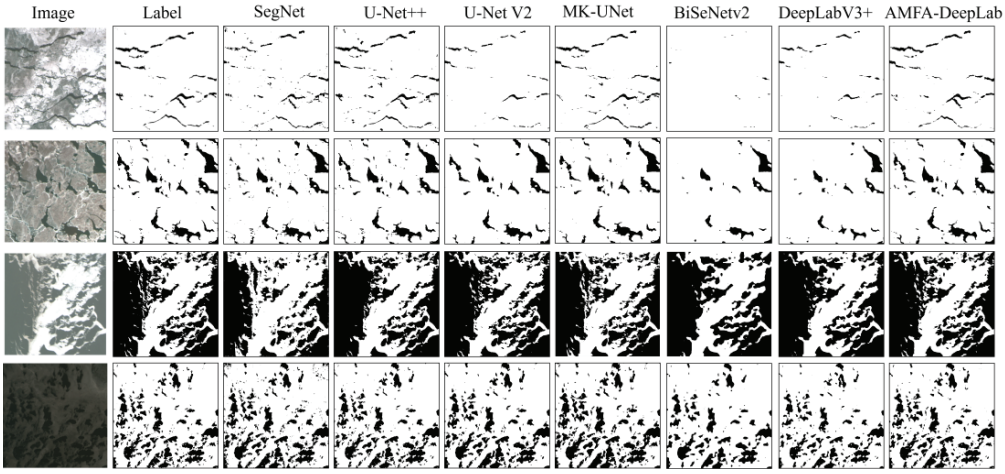


Figure 8. Extraction results based on different segmentation networks.

As can be discerned from the scene of narrow ice cracks in the first row, the results based on the SegNet and DeepLabV3+ models exhibit severe breaking phenomena, failing to maintain the continuity of ice cracks, so that long cracks are incorrectly segmented into unconnected short line segments. It is worth noting that the lightweight BiSeNetV2 almost completely missed the detection of fine ice cracks due to its limited feature capacity. Although the results based on U-Net++ restore the connectivity to a certain extent, detection of tiny parts of the crack tips are still missed. Comparatively, AMFA-DeepLab can accurately capture fine cracks with widths of only several pixels, with generated segmentation masks highly coinciding with the labels, demonstrating the advantage of the AMFA module with respect to maintaining high-frequency linear features.

In the scene of dense, fragmented sea ice blocks in the second row, there are numerous ice blocks with narrow gaps between them, and there are tiny, crushed ice residues scattered in the background, which can easily lead to adhesion or noise in the segmentation results. In this scene, owing to the severe interference from background fragments, SegNet outputs evident salt-and-pepper noise and fails to clearly identify the ice block boundaries. DeepLabV3+ and BiSeNetV2 tend to wrongly merge adjacent ice blocks, losing the narrow water channel features between ice blocks. The AMFA-DeepLab proposed in this study not only effectively filters off tiny interference in the background but also accurately separates closely adjacent independent ice blocks, clearly restoring the complete contour of each fragmented ice block without adhesion.

The third row corresponds to a scene of complex drift ice, with relatively complex textures. In this scene, the SegNet model lacks effective aggregation of contextual information, resulting in loss of many drift ice textures. U-Net++, U-NetV2, and MK-UNet reserve rough contours, but significant edge overlapping occurs in processing gradient details in the drift ice, easily resulting in adhesion. AMFA-DeepLab successfully captures the dynamic textures specific to drift ice, clearly restoring the fine fragmentation features in the surface of drift ice. It successfully reserves the complex texture structures in sea ice and accurately restores the gradient details at ice–water boundaries, reflecting the powerful analytical ability of the AMFA module regarding texture features.

The fourth row is a scene of coexisting nearshore land and sea ice. It can be seen that missed detections occur in the results based on SegNet, U-Net++, DeepLabV3+, and BiSeNetV2, with evident serration appearing. The edge curve generated by the model proposed in this study is smooth and quite close to the true boundary; in particular, extremely high positioning precision is exhibited at concave–convex transition locations on land.

https://doi.org/10.3390/rs18050783

In summary, the AMFA-DeepLab proposed in this study is significantly better than SegNet, U-Net++, U-NetV2, BiSeNetV2, MK-Unet, and DeepLabV3+ in terms of preserving the continuity of fine cracks, suppressing background noise, and restoring the complex texture details. These results fully verify that the multi statistics fusion strategy of AMFA module, which can effectively overcome the inherent semantic fuzziness in complex sea ice region, has strong robustness.

3.3. Results of the Generalization Tests

To further explore the adaptability of the model to complex sea ice texture across different years or during the transition to the melting season, a full-scale GF-1 WFV image obtained on 2 February 2017 was used for generalization test. This data was strictly excluded from the model training stage, as a challenging case.

In Figure 9, the original GF-1 WFV image shows the complex sea ice distribution, especially in the bottom of the sea ice region; it presents a complex filiform structure formed by drift ice, which is very difficult to retain via traditional segmentation models. The ice identification result confirms that the AMFA-DeepLab still has strong generalization ability. The result not only outlines the overall sea ice range, but also accurately captures the fine filament structure at the bottom, effectively distinguishing it from the open sea water area. Those results confirm that the model has learned robust feature representation, can adapt to complex sea conditions on other years, and is suitable for operational monitoring requirements and tasks in Liaodong Bay in the Bohai Sea, China.

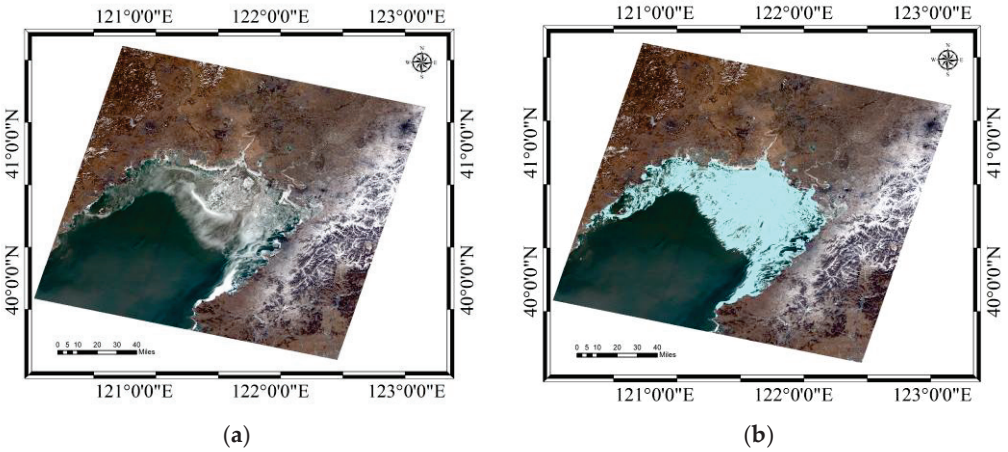


Figure 9. The GF-1 WFV image on 2 February 2017 (**left**) and the ice identification result (**right**). (**a**) The original GF-1 WFV image shows the complex distribution of drift ice and the filiform texture at the bottom. (**b**) Overlay of (light blue is the predicted sea ice mask).

4. Discussion

The proposed model for sea ice segmentation based on an improved lightweight DeepLabV3+ adaptive multi-statistic fusion attention architecture demonstrates improvements in model training efficiency, extraction accuracy, and noise resistance robustness compared to those of traditional models. By integrating the MobileNetV2 and AMFA modules, the model achieves lightweight training and captures more nuanced multiscale features, effectively overcoming the segmentation challenges posed by blurred texture features and weak boundary details in sea ice identification.

Unlike many studies that introduce complex attention mechanisms resulting in significantly increased computational load, this research successfully achieves a dual optimization of performance and efficiency by reconstructing the backbone network into a lightweight MobileNetV2. As shown in Table 1, the model's training time is substantially reduced to 4.42 h, and the number of parameters decreases by approximately 90% compared to the

https://doi.org/10.3390/rs18050783

original DeepLabV3+ model. Importantly, its inference speed reaches 281.76 FPS, which is significantly better than the U-Net++ (107.91 FPS) and SegNet (121.23 FPS). Although the inference speed of BiSeNetv2 is faster, its IOU (87.80%) is much lower than that of AMFA-DeepLab (92.15%). Therefore, AMFA-DeepLab achieves high inference efficiency while maintaining high accuracy, offering far superior computational efficiency and high accuracy in segmentation, better meeting the real-time requirements for emergency sea ice disaster monitoring.

The notable advantage of the proposed network model lies in its capability for fine-grained segmentation, which is particularly crucial for identifying irregularly shaped broken ice floes and narrow ice cracks. Traditional image segmentation processing approaches are often susceptible to interference from speckle noise. Some network models, including the U-Net series and DeepLabV3+, perform adequately in capturing overall contours and tend to produce fractures or adhesion when handling subtle textures. In contrast, the proposed AMFA-DeepLab incorporates a noise-resistant mechanism through the global median pooling within the AMFA module, which effectively suppresses false activations in non-ice regions, thereby significantly enhancing segmentation accuracy in edge areas. As shown in Figure 7, the architectural optimization has led to more accurate identification results. Particularly in complex drifting ice scenarios, the model successfully reduces false positives and clearly restores the transitional details at the ice–water interface. As shown in Figure 9, the model can still have the ability to capture the fine filament structure from some images taken in other years or during the melting period. The dynamic channel mixing mechanism of AMFA facilitates the network's adept utilization of global contextual information while achieving adaptive feature recalibration through the fusion of average, max, and median statistics. This improvement not only reduces false detections but also substantially improves segmentation performance in challenging edge regions.

It is still a challenge to achieve appropriate segmentation performance across different seasons, specific regions, and complex environmental conditions. During the warm season, melting ice may significantly change the spectral reflectance of sea ice, leading to confusion with open water. While in the cold season, the heavy snow covering sea ice may smoothen the image texture details, posing a challenge to the model that depends on texture features. In addition, the current training dataset only covers the Liaodong Bay Area in the Bohai Sea, China, where the ice is mainly one-year ice, and it is less affected by snow cover. In contrast, polar environments usually contain complex multi-year sea ice, and some deep covered snow or some melting ice pools. The AMFA-DeepLab is trained based on the specific sea ice texture features on Liaodong Bay, and it may encounter domain offset problems when directly applied to polar regions. Therefore, it still needs to verify the generalization ability in polar regions with different ice types through transfer learning or domain adaptation technology.

However, the current model still has some limitations, primarily constrained by its reliance on optical remote sensing image. This study solely employs visual RGB images from Chinese GF-1 satellite WFV sensor and does not incorporate infrared imagery or integrate other spectral information. Consequently, the model may miss some texture information that is more sensitive to thermal or multiple spectra, which to some extent affects the model's generalization ability. Additionally, The AMFA module exhibits insufficient capability for capturing the features of newly formed and melting ice image, leading to challenges in segmenting certain thin ice areas. Future research will focus on multimodal data fusion and dynamic temporal analysis. On one hand, to address the inherent drawback of optical imagery, the thermal infrared images and some spectral information will be introduced to expand the scale and diversity of the dataset and enhance model generalization. Furthermore, Synthetic Aperture Radar (SAR) data will be incorporated to construct

https://doi.org/10.3390/rs18050783

an optical-infrared-radar feature fusion network, thereby strengthening the capability for all-weather sea ice segmentation and classification. On the other hand, considering that the formation and evolution of sea ice constitute a dynamic and evolving process, the integration of time-series remote sensing images with long-term sequence models can be adopted to enhance sea ice segmentation or identification. This future study will not only improve the sea ice monitoring, but will also enable tracking and prediction of sea ice, thereby providing more proactive and valuable decision support for disaster prevention in fields such as shipping, fisheries, and offshore operations.

5. Conclusions

Sea ice monitoring based on Chinese GF-1 WFV images suffers from large computational redundancy, weak antinoise ability, and easily missed detection of fine features. To address these problems, we constructed a modified lightweight AMFA-DeepLab segmentation network model. Through theoretical analysis, model construction, ablation experiments comparative experiments, and generalization tests, the following main conclusions can be drawn: (1) Replacing the Xception network with the MobileNetV2 backbone network greatly decreases the number of model parameters—from 54.71 million to 5.85 million, which is only 10.7% of those of the DeepLabV3+ model, and the training time is shortened by nearly 20%. Experimental results show that the linear bottleneck layer and inverted residuals of MobileNetV2 can even more efficiently extract sea ice features with the inference speed as high as 281.76 FPS without negatively impacting segmentation precision but with significant reduction in demand for computing capability, thereby successfully solving the difficult problem of timeliness in monitoring large-area sea ice. (2) For the speckle noise and weak texture feature specific to sea ice images, we proposed an AMFA module, which, through introducing global median pooling and dynamic weight distribution, effectively overcomes the limitations of traditional single pooling strategy. The ablation experiment and heatmap visualization analysis consistently demonstrate that this model significantly suppresses incorrect activation of nearshore land and seawater background (false detection) and greatly improves the response to dense fragmented ice blocks and narrow ice gaps (solving the problem of missed detection). The final IoU of the model reaches 92.15%, and it is the best of all of the models used for comparison. The generalization tests show that the AMFA-DeepLab has good cross-temporal and spatial robustness. Even for the complex drift ice scene during the melting period in 2017, the model can still accurately capture the filiform texture structure of sea ice, proving that it can adapt to inter-annual changes and some environmental disturbances. The AMFA-DeepLab designed in this study achieves high-precision segmentation of sea ice under complex marine conditions while maintaining extremely low model complexity. Not only is it better than the traditional SegNet and U-Net series models, but it achieves a balance between model computational efficiency and segmentation precision, overcoming the limitation that a lightweight network has in maintaining high precision. This method has strong robustness capability, and it can meet the demand for rapid emergency response monitoring of sea ice disasters based on Chinese GaoFen-1 satellite, thus being of important engineering application value. In future work, we will focus on improving the generalizability of the model by expanding the scale and diversity of the training datasets, which include different ice images on polar region and enhancing the capability for continuous dynamic monitoring of sea ice.

Author Contributions: Conceptualization, Z.H. and D.P.; methodology, Z.H.; software, X.L. and J.C.; validation, Z.H., X.L., and Q.Z.; formal analysis, Z.H.; investigation, X.L. and Y.L.; resources, Q.Z.; data curation, X.L. and Q.Z.; writing—original draft preparation, Z.H. and X.L.; writing—review and editing, Z.H., Z.M., and D.P.; visualization, X.L.; supervision, D.P.; project administration, Z.H. and

https://doi.org/10.3390/rs18050783

D.P.; funding acquisition, Z.H. and Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Joint Funds of the Zhejiang Provincial Natural Science Foundation of China Under Grant No. LZJMZ24D060001, and Major Innovation Project of Science, Education and Industry Integration Pilot Project of Qilu University of Technology (Shandong Academy of Sciences) Under Grant No. 2025ZDYS01.

Data Availability Statement: The GF-1 WFV image in this study can be downloaded from China Platform of Earth Observation System (https://www.cpeos.org.cn/home/#/, accessed on 9 April 2025). SegNet can be downloaded from (https://github.com/alexgkendall/SegNet-Tutorial, accessed on 8 May 2025). U-Net++ can be downloaded from (https://github.com/MrGiovanni/UNetPlusPlus, accessed on 8 May 2025). U-NetV2 can be downloaded from (https://github.com/yaoppeng/U-Net_v2, accessed on 8 May 2025). DeepLabV3+ can be downloaded from (https://github.com/chenxi116/DeepLabv3.pytorch, accessed on 8 May 2025). BiSeNetv2 can be downloaded from (https://github.com/CoinCheung/BiSeNet, accessed on 15 January 2026). MK-UNet can be downloaded from (https://github.com/SLDGroup/MK-UNet, accessed on 15 January 2026).

Acknowledgments: We would like to thank the China Platform of Earth Observation System (CPEOS), which provided the GaoFen-1 satellite data. The authors would also like to thank the editors and reviewers for their valuable comments.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AMFA	Adaptive Multistatistic Fusion Attentione
AMFA-DeepLab	AMFA module using DeepLabV3+ as the base architecture
GF-1	GaoFen-1
SAR	Synthetic Aperture Radar
WFV	wide field of view
GF-3	GaoFen-3
ASPP	Atrous Spatial Pyramid Pooling
MLP	Multi-Layer Perceptron
1D	one-dimensional
BCE	binary cross entropy
TP	true positive
FP	false positive
FN	false negative
P	precision
R	recall
F1	F1-score
IoU	intersection over union
FPS	frames per second

References

1. Geng, S.S.; Han, C.H.; Yang, Y.; Zheng, B. An analysis on the economic losses of sea ice disasters and defense countermeasures in bohai sea. *Mar. Econ.* **2023**, *13*, 101–107. [CrossRef]
2. Liu, W.P.; Yan, D.W.; Peng, Z.K.; Xie, M.H.; Sun, Y.L. Vessel safety navigation under the influence of antarctic sea ice. *J. Mar. Sci. Eng.* **2025**, *13*, 1267. [CrossRef]
3. Wang, C.L.; Wu, J.D.; He, X.; Ye, M.Q.; Liu, Y. Quantifying the spatial ripple effect of the Bohai Sea ice disaster in the winter of 2009/2010 in 31 provinces of China. *Geomat. Nat. Hazards Risk* **2018**, *9*, 986–1005. [CrossRef]
4. Liu, H.Y.; Guo, H.D.; Zhang, L. Sea ice classification using dual polarization SAR data. *IOP Conf. Ser. Earth Environ. Sci.* **2014**, *17*, 012115. [CrossRef]

https://doi.org/10.3390/rs18050783

5. Zou, J.H.; Zeng, T.; Guo, M.H.; Cui, S.X. The study on an Antarctic sea ice identification algorithm of the HY-2A microwave scatterometer data. *Acta Oceanol. Sin.* **2016**, *35*, 74–79. [CrossRef]
6. Li, X.N.; Zhang, J.; Dai, Y.S.; Zhang, X. Research on the enhanced performance of texture feature for sea ice drift monitoring based on gray level co-occurrence matrices. *Mar. Sci.* **2018**, *42*, 9–17. [CrossRef]
7. Zhou, Y.; Kuang, D.B.; Gong, C.L.; Hu, Y.; Fang, S.H.; Zhang, Y.; Peng, Y. A method to extract parameters of Arctic sea ice from FY-3/MERSI imagery. *J. Infrared Millim. Waves* **2017**, *36*, 41–48+126–127. [CrossRef]
8. Li, T.; Wu, D.; Han, R.; Xia, J.Y.; Ren, Y.J. A sea ice recognition algorithm in bohai based on random forest. *Comput. Mater. Contin.* **2022**, *73*, 3721–3739. [CrossRef]
9. Lia, L.; Suo, Z.Y.; Shi, L.J.; Wang, Q.; Jiao, J.N.; Tang, J.; Lu, Y.C. Identification and estimation of sea ice in Liaodong Bay using HY-1C/D satellite images. *Natl. Remote Sens. Bull.* **2025**, *29*, 897–909. [CrossRef]
10. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015*; IEEE: Piscataway, NJ, USA, 2015.
11. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015*; Navab, N., Hornegger, J., Wells, W., Frangi, A., Eds.; Springer: Cham, Switzerland, 2015.
12. Xu, H.; Ren, Y.B. Detecting sea ice of Bohai Sea using SAR images based on a hybrid loss U-Net model. *Acta Oceanol. Sin.* **2021**, *43*, 157–170. [CrossRef]
13. Zhang, T.Y.; Yang, Y.; Shokr, M.; Mi, C.L.; Li, X.M.; Cheng, X.; Hui, F.M. Deep learning based sea ice classification with Gaofen-3 fully polarimetric SAR data. *Remote Sens.* **2021**, *13*, 1452. [CrossRef]
14. Dai, Y.; Li, X.-M.; Yuan, H. Pan-Arctic winter sea ice classification using Sentinel-1 dual-polarized SAR images. *Remote Sens. Environ.* **2026**, *333*, 115140. [CrossRef]
15. Sun, S.C.; Wang, Z.Y.; Li, Z.J.; Zhang, B.J.; Tian, K.; Zhao, X.Y. An extraction method for sea ice based on improved DeepLabV3+model: Taking the Arctic Greenland Sea as an example. *Acta Oceanol. Sin.* **2024**, *46*, 131–142. [CrossRef]
16. Han, Y.; Gao, Y.; Zhang, Y.; Wang, J.; Yang, S. Hyperspectral sea ice image classification based on the spectral-spatial-joint feature with deep learning. *Remote Sens.* **2019**, *11*, 2170. [CrossRef]
17. Han, Y.; Cui, P.; Zhang, Y.; Zhou, R.; Yang, S.; Wang, J. Remote sensing sea ice image classification based on multilevel feature fusion and residual network. *Math. Probl. Eng.* **2021**, *2021*, 9928351. [CrossRef]
18. Zhou, L.; Xu, R.X.; Bian, J.Y.; Ding, S.F.; Han, S.; Skjetne, R. Optimized recognition algorithm for remotely sensed sea ice in polar ship path planning. *Remote Sens.* **2025**, *17*, 3359. [CrossRef]
19. Zheng, B.; Shi, L.J.; Zou, B.; Ren, P.; Zeng, T.; Sun, X.Y.; Zhang, C.H. Research on Sentinel-1 SAR sea ice detection method in Liaodong Bay based on AUNet++. *Acta Oceanol. Sin.* **2024**, *46*, 117–128. [CrossRef]
20. Chen, L.C.; Zhu, Y.K.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the Computer Vision—ECCV 2018, Munich, Germany, 8–14 September 2018*; Lecture Notes in Computer Science; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Springer: Cham, Switzerland, 2018.
21. Chen, S.; Wei, X.; Zheng, W. ASA-DRNet: An Improved Deeplabv3+ Framework for SAR Image Segmentation. *Electronics* **2023**, *12*, 1300. [CrossRef]
22. Yang, T.; Wei, J.; Xiao, Y.; Wang, S.; Tan, J.; Niu, Y.; Duan, X.; Pan, F.; Pu, H. LT-DeepLab: An Improved DeepLabV3+ Cross-Scale Segmentation Algorithm for Zanthoxylum bungeanum Maxim Leaf-Trunk Diseases in Real-World Environments. *Front. Plant Sci.* **2024**, *15*, 1423238. [CrossRef]
23. Feng, X.; Wei, C.; Xue, X.; Zhang, Q.; Liu, X. RST-DeepLabv3+:Multi-Scale Attention for Tailings Pond Identiffcation with DeepLab. *Remote Sens.* **2025**, *17*, 411. [CrossRef]
24. Shaheed, K.; Mao, A.; Qureshi, I.; Kumar, M.; Hussain, S.; Ullah, I.; Zhang, X. DS-CNN: A pre-trained Xception model based on depth-wise separable convolutional neural network for finger vein recognition. *Expert Syst. Appl.* **2022**, *191*, 116288. [CrossRef]
25. Mekruksavanich, S.; Phaphan, W.; Jitpattanakul, A. Squeeze-and-Excitation Hybrid Network for Biometric Identification via Photoplethysmography. In *Proceedings of the 2024 Research, Invention, and Innovation Congress: Innovative Electricals and Electronics (RI2C), Bangkok, Thailand, 8–9 August 2024*; IEEE: Piscataway, NJ, USA, 2024.
26. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018*; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Springer: Cham, Switzerland, 2018.
27. Milletari, F.; Navab, N.; Ahmadi, S.-A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016*; IEEE: Piscataway, NJ, USA, 2016.
28. Loshchilov, I.; Hutter, F. Sgdr: Stochastic gradient descent with warm restarts. *arXiv* **2016**, arXiv:1608.03983. [CrossRef]
29. Everingham, M.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. [CrossRef]

https://doi.org/10.3390/rs18050783

30. Zhou, Z.; Rahman Siddiquee, M.M.; Tajbakhsh, N.; Liang, J. Unet++: A nested u-net architecture for medical image segmentation. In *Proceedings of the International Workshop on Deep Learning in Medical Image Analysis, Granada, Spain, 20 September 2018*; Springer: Cham, Switzerland, 2018.
31. Peng, Y.; Chen, D.Z.; Sonka, M. U-net v2: Rethinking the skip connections of u-net for medical image segmentation. In *Proceedings of the 2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI), Houston, TX, USA, 14–17 April 2025*; IEEE: Piscataway, NJ, USA, 2025.
32. Rahman, M.M.; Marculescu, R. MK-UNet: Multi-kernel lightweight CNN for medical image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Honolulu, HI, USA, 19–23 October 2025*; IEEE: Piscataway, NJ, USA, 2025; pp. 1042–1051.
33. Yu, C.; Gao, C.; Wang, J.; Yu, S.; Shen, C.; Sang, N. BiSeNet V2: Bilateral Network with Guided Aggregation for Real-Time Semantic Segmentation. *Int. J. Comput. Vis.* **2021**, *129*, 3051–3068. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

https://doi.org/10.3390/rs18050783

GeoHazards

MDPI

Article

A Context-Aware Flood Warning Framework Integrating Ensemble Learning and LLMs

Adnan Ahmed Abi Sen [1,*], Fares Hamad Aljohani [2], Nour Mahmoud Bahbouh [3], Adel Ben Mnaouer [1], Omar Tayan [1] and Ahmad. B. Alkhodre [4]

[1] Hussein ElSayyed Research and Innovation Center, Deanship of Scientific Research and Graduate Studies, University of Prince Mugrin, Madinah 42241, Saudi Arabia; a.mnaouer@upm.edu.sa (A.B.M.); o.tayan@upm.edu.sa (O.T.)
[2] Department of Information Systems, Northern Border University, Rafha 91911, Saudi Arabia; fnahezaljohani@stu.kau.edu.sa
[3] Department of Information and Communication Sciences, Granada University, 18071 Granada, Spain; nourmahmoud@correo.ugr.es
[4] Department of Information Technology, Islamic University, Madinah 42351, Saudi Arabia; aalkhodre@iu.edu.sa
* Correspondence: a.ahmed@upm.edu.sa

Abstract

Smart cities require effective disaster management (like flooding, solar storms, sandstorms, or hurricanes), as it directly impacts people's lives. The key challenges of disaster management are timely detection and effective notification during the crisis. This research presents a smart multi-layer framework for notification classification and management before and during flooding disasters. The framework includes an early detection module as the main phase in the alerting process. This step depends on an Ensemble Learning (EL) model based on a triad of the three best selected models (Deep Learning (DL), Random Forest (RF), and K-nearest Neighbor (KNN)) to analyze data collected continuously from the Internet of Things (IoT) layer. In the boosting phase, the framework utilizes Large Language Models (LLMs) with DL to analyze social textual crowdsourcing data. The results will enable the framework to identify the most affected areas during a flood. The framework adds a fog computing layer alongside a cloud layer to enable instantaneous processing of user responses and generate specialized alerts based on contextual factors such as location, time, risk level, alert type, and user characteristics. Through testing and implementation, the proposed algorithms demonstrated an accuracy rate of over 98% in detecting threats using a dataset of real, collected weather and flooding data. Additionally, the framework proposes a centralized control panel and a design of a smartphone application that offers essential services and facilitates communication among managed civil defense teams, citizens, and volunteers.

Keywords: alert; notification; crowdsourcing; IoT; fog computing; AI; hazard; disaster

1. Introduction

Disaster management is vital for smart cities because it affects people's lives and connects to systems like health, transportation, security, and safety. Natural disasters include events like landslides, earthquakes, wildfires, tornadoes, floods, tsunamis, and pandemics. Early warnings are key to reducing the damage caused by disasters. Managing these events is challenging and urgent, as it relates to weather, climate change, and water resources on Earth [1].

Floods are highly destructive and cause major damage. Tomar et al. [2] reported that between 2000 and 2014, floods made up 39.26% of global natural disasters, causing USD 397.3 billion in losses. EM-DAT data shows that Asia is the hardest-hit region by floods and storms. Much of this damage is linked to human actions like deforestation, urbanization, and poor planning. Climate change has also worsened flooding, leading to heavier rainfall [3]. Floods cause major losses and are expensive to manage [4]. Their environmental impact is deep and complex. They immediately result in property damage and loss of life. Floods can also lead to livestock deaths, crop destruction, and a rise in water-borne diseases [5].

Floods also slow down economic growth because of the high costs of relief and recovery efforts. This can harm infrastructure projects and other development plans, possibly weakening the regional economy (see Figure 1) [6]. It is crucial to take action to reduce these negative effects. Accurately identifying flood hazard zones requires considering factors like elevation, rainfall patterns, land use, flow accumulation, and slope. These factors interact and create challenges. Technologies like the Internet of Things (IoT) and machine learning (ML) help predict floods but have limitations.



Figure 1. Jeddah's disaster after flooding in 2023.

For example, different areas within the same city may require different responses [7,8]. The most vulnerable flood zones are determined by many key factors. In addition to elevation, rainfall intensity, land use, flow accumulation, and slope, an important factor is the watershed lag time, which is a critical hydrological factor influencing flood response speed. Lag time reflects the temporal delay between peak rainfall and peak runoff discharge. The level of risk varies across city areas, as these factors affect flood hazards differently [8,9].

In flood risk management, most research methods have used traditional physical techniques and machine learning for prediction. However, these methods can be imprecise, as flood risk varies by region and even within neighborhoods, based on local geography. As a result, different areas within the same city may have different levels of flood vulnerability. For example, low-lying and densely populated areas are at higher risk of flooding and losses. To improve prediction accuracy, algorithms need to use multiple data sources [9].

Accurate and reliable data are essential for effective flood management models. Real-time data is mainly provided through the Internet of Things (IoT), which uses wireless sensor networks (WSNs) to monitor environmental conditions like temperature, pressure, water levels, and more across various locations. IoT also employs radio frequency identification (RFID) technology to track and connect objects based on their locations [10,11]. Notably, related studies can also be found in the literature that present other factors related

https://doi.org/10.3390/geohazards7010035

to climate, disasters, and pollution monitoring in a smart city context that employ ML, cloud computing, and IoT [12].

Crowdsourced data from mobile devices and social media can improve accuracy and situational awareness when combined with the IoT. For dynamic events like floods, human input ensures the reliability of crowdsourced data. These models rely on large amounts of rapidly generated data to tackle flood-related challenges. In recent years, crowdsourcing has become a key data source, as smartphones and social media provide fast, valuable, real-time information [13,14]. Crowd-sourced data can often be more useful than IoT data because it relies on human senses and judgment to filter and interpret information before it is sent. For example, analyzing camera data automatically during a flood can be harder and less accurate than having people describe what they see, followed by machine analysis [14].

Textual data from crowdsourcing, like social media, enables highly accurate and context-aware text analysis. Large Language Models (LLMs) demonstrate remarkable capabilities in understanding the semantic meaning, sentiment, and implicit intent of unstructured textual data. In disaster management, such as floods, LLMs can be very effective in analyzing emergency calls on social media, classifying them into meaningful classes, and extracting the most critical situations and areas to take the right secure action. LLM-based systems can interpret local expressions and event-specific terminology and provide a real-time situational awareness for decision-making [15].

LLMs have become an entity that can accept and reject, say and decide, and invent and recommend, not only an algorithmic system. They can understand polarization through human interactions and extract sentiment and topics from social networks. However, there are still doubts and threats about the responsibility of these models when they decide without supervision, in addition to the threat of hallucinations and the bias in their information [16–18].

Combining IoT-based data collection with crowdsourced data is especially helpful in situations like disaster management, traffic monitoring, or events, where people can share useful information through their observations. This paper proposes an intelligent solution for early flood detection that focuses on reliability and accuracy by integrating IoT-based sensing and crowdsourcing modules. The contributions of this paper include:

- A flood detection framework that combines IoT-based and crowdsourced data to provide efficient responses.
- A fog-based architecture to reduce the response delay and provide context-aware notifications to stakeholders.
- A novel threat level classification Ensemble Learning (EL)-based algorithm that ensures higher reliability of the decision.
- An LLM base with a DL model to analyze textual data during a disaster.
- A mockup mobile application for volunteers and a dashboard for civil defense officers.

This paper relies on two types of data to provide comprehensive coverage compared to other solutions; the types are:

- Data from IoT devices (WSNs): These are used continuously during the first classification phase (Algorithm 1). The system here is based on ten years of meteorological data obtained from the weather authority in the city of Jeddah in Saudi Arabia.
- Crowdsourced data: This type of data is expected to come from volunteer users who have the application, in the form of distress calls or textual information. These data are used within Method 2, but only when the risk index from Algorithm 1 exceeds a certain threshold. (More details were presented in Section 4).

https://doi.org/10.3390/geohazards7010035

Algorithm 1: EL algorithm for training and testing the dataset

Require: Data in the cloud
Ensure: Data processing and splitting the dataset into training and testing
1: **for** all data in the cloud **do**
2: Process outlier values by replacing them with the median
3: **for** each record containing empty values, **do**
4: Delete records
5: Apply normalization to numerical data
6: Use the Synthetic Minority Over-Sampling Technique (SMOTE) to address the imbalance
7: Split the dataset into training and testing
8: Train the dataset on the training part
8: **If performance** metrics are the best for a model candidate, **then**
9: Select the winning model hyperparameters
10: **else**
11: Select other values for hyperparameters
12: **end if**
13: **end for**
14: **end for**
16: **for** testing phase in the fog node (for a specific region), **do**
17: Read data from the sensor layer
18: Process data and detect outliers by matching values across multiple sensors
19: Configure data and apply normalization.
20: Apply the three **selected machine learning models (as the chosen EL set).**
21: Choose the majority decision (minimum 2 models from 3 classified data as 1 or 0).
23: **if** the choice is 1 (threat present), **then**
24: Increase the threat index by 1
25: **else**
26: Keep the same value
27: if the threat level exceeds a certain threshold, **then**
28: Send immediate warning alerts according to the region
29: **else**
30: Keep monitoring
31: **end if**
32: end if
33: **end for**

The rest of the paper is structured as follows: Section 2 reviews related work in disaster management. Section 3 outlines the proposed approach, framework, tools, and algorithms. Section 4 presents the results and insights, while Section 5 concludes and summarizes key findings and contributions.

2. Related Work

Within this section, an extensive exploration of the literature about flood risk management models is undertaken. Specifically, this encompasses a detailed examination of two primary models: The Smart Model and the Physical Model. The discussion delves into the intricacies and applications of these models within the context of flood risk management.

Furthermore, an analysis of tweets as a novel approach to understanding and evaluating flooding risk is presented, elucidating the significance and potential insights gleaned from social media data in this domain.

https://doi.org/10.3390/geohazards7010035

2.1. Models Used in Flood Risk Management

Floods pose immense threats to lives and property, impacting the environment, social fabric, and economies globally. Recent years have underscored an urgent demand for precise flood modeling to aid decision-makers and communities in responding to, managing, and alleviating flood impacts. Among the most formidable obstacles encountered by experts is the intricate task of forecasting floods accurately. The precision of predictions plays a pivotal role in devising effective strategies to mitigate the severity of floods and minimize their potential devastation to lives and properties [19].

The most widely used models in flood risk management are physical modeling and intelligent data-based modeling. Physical modeling is based on hydraulic experiments, through which predictive maps are generated showing areas that may be flooded. On the other hand, machine learning approaches rely on data and analysis [20]. Table 1 summarizes methods for dealing with flooding.

Table 1. Summary of flooding management methods.

Ref.	Methodology	Objective	Smart	Physical
[21]	GIS spatial statistics	Assess and analyze flooding risk in Chongqing, China	No	Yes
[22]	WSNs with IoT	Investigate the roles of IoT and WSNs in disaster management		
[23]	GIS and hydraulic	Study flood risk for different land uses in Surma, Bangladesh	Yes	No
[24]	Hydrological modeling	Identify flood-prone urban areas		
[25]	ANN and LR	Demonstrate that machine learning techniques can be used to accurately map and predict flood-prone areas and to develop flood mitigation plans and policies	No	Yes
[26]	Spatial analysis	Identify and map the city of Jeddah's flood zones to minimize their susceptibility and include them in flood risk prevention and mitigation methods	No	Yes
[27]	RFR model	Identify high-risk areas	Yes	No
[28]	IoT and ML models	Determine the level of risk in each area of a city	No	Yes
[29]	ML models	Use ML models to predict the occurrence of urban pluvial flooding	No	Yes

Intelligent data-driven modeling and the Physical Model have risen to prominence as leading methodologies within the realm of flood risk management. These approaches reflect various perspectives for solving the problems brought on by flood disasters. The Smart Model improves early warning systems by providing rapid and precise flood predictions using real-time data from the IoT and advanced analytics as text analysis. The Physical Model, on the other hand, depends on accurate simulations of flood scenarios, providing insightful information about hydraulic processes and assisting in the comprehension of flood behavior [30].

2.2. Text Analysis in Flooding Risk

Social media data can help crisis managers and responders meet their data needs by providing valuable information on the social aspects, impacts, and flexibility of cities in the face of a natural disaster [31]. Using Twitter data for flood risk analysis is a promising new field of study. This is because Twitter data can provide real-time information about

https://doi.org/10.3390/geohazards7010035

flood events, which can be used to improve early warning systems and assist emergency responders in flood event management [32].

However, some issues must be addressed, such as the need to develop more precise methods for extracting and analyzing Twitter data. Text analysis entails using natural language processing techniques to extract information from Twitter posts, such as the flood's location, severity, and impact on people and infrastructure [33].

According to the findings of some studies [34,35] in the field of tweet analytics in flood risk management, Twitter data can be used to predict the location of floods with a high degree of accuracy. Twitter data can also be used to identify flood-prone areas before they occur [36]. The text of the tweet contains helpful and trustworthy information about damaged main roads and streets. This information could help with emergency response coordination and resource allocation [37].

Crowdsourcing is a model that uses people's intelligence via online human input to achieve specific organizational goals [38]. Crowdsourcing, according to a study conducted by Tripathy et al., provides a viable source of reliable information on floods and waterlogging in Mumbai. Crowdsourced data can detect hotspots and has the potential to generate real-time monitoring. This information can then be used to create a flood forecasting framework. When fine-resolution observed datasets are unavailable, crowdsourced information can be extremely useful. These datasets can be used to create a robust, modern, and decision-making system that will allow for more precise and effective decisions in the event of a disaster [39].

Although there is interest from researchers in studies on social media platform-based crowdsourcing in disaster management, smartphone sensor-based crowdsensing in disaster incidents appears to still be in need of more research. Table 2 summarizes the algorithms and methods for traditional text analysis based on ML and text mining (TM), as well as a few new methods that depend on LLMs. The main steps and functions of traditional methods include tokenizing, stemming, lemmatizing, removing stop words, weighting, and training models, while new models rely on LLMs to understand semantics and cluster topics or tag data like tweets automatically. Semantics can provide more information about the type of threat, severity, effects, location, and sentiment.

Table 2. Summary of using text analysis for flooding management.

Ref.	Objectives	Model	Results	Limitations
[34]	Use an automatic tweet parsing system, effectively use social media in locating users asking for help during a disaster	Markov model	Develop a TM algorithm to detect flood-related tweets in English and Hindi, as well as classify these tweets into high and low priority to identify those that require attention.	Some tweets are misclassified by the proposed system and can be studied by researchers to determine the reasons for such misclassification.
[40]	Describe the flood alert situation using only tweet messages and investigate if the informative potential of such data is also demonstrated	Naïve Bayes	Twitter messages contain valuable flood spatial information, according to text analysis techniques.	Because of the complexity of some language structures, which contain many special characters, some languages will be difficult to implement in such an environment.

https://doi.org/10.3390/geohazards7010035

Table 2. *Cont.*

Ref.	Objectives	Model	Results	Limitations
[41]	Investigate accurate classification for short, informal (colloquial) Arabic tweets	R tool with SVM	Using colloquial Arabic text as a dataset, investigated a variety of text classification techniques.	There is a reliability issue because it relies on only one model with low accuracy and a small dataset.
[42]	Present the findings of an analysis using an innovative methodology and use the 2010–2011 Southeast Queensland Floods as a case study to demonstrate how disaster severity can be assessed using tweets	Decision tree	The research presented contributes to a better understanding of the systematic use of volunteer crowdsourced data to improve disaster management practices.	The inequality of geo-located tweets is a critical constraint.
[43]	Build a model to categorize tweets to better organize rescue and relief operations and save lives	Neural Network	The paper compares several conventional machines and Deep Learning techniques.	For the classification task, only English-language tweets were used, whereas during disasters, users posted in their regional languages.
[44]	Make a map that characterizes the social impacts of the major flood event in Kerala in 2018	Manual inspection	Flood impact maps derived from Telegram and Twitter.	On Twitter, relevant data is mixed in with larger amounts of irrelevant data, which means the data needs more filtering steps.
[45]	Use LLM-based multi-agent models for multiclassification of tweets	LLM	Better results of classification compared to traditional LM models.	Stability of performance and responsibility of autoclassification.
[46]	Utilize LLMs and ML to classify social media of disasters	LLM + ML	Detect the type of threat and its severity, with additional information reducing the time needed for tagging training data.	Depend only on textual and multi-topic data.

Although LLMs provide more information and summarize the time required for text processing, depending solely on LLMs (which are unsupervised models) can lead to hallucinations and bias, which are very dangerous in critical systems like disaster management [45,46]. Although these methods (traditional and new ones) make important contributions to text classification and information retrieval in many applications, they still face some drawbacks (depend on many manually designed features with sequence processing steps, do not utilize cumulative training and pre-trained models, such as transfer learning models or adaptive models like DL, and suffer from hallucinations and bias in LLM cases).

Based on the previous discussion, this research merged LLMs and DL to get the benefits of LLMs and avoid their shortcomings. LLMs save training time and data and increase semantic meaning understanding, while DL increases accuracy and adaptability and avoids the hallucinations and biases of LLMs. So, the semi-supervised proposed model will achieve a key transformation in the domain of natural language processing

https://doi.org/10.3390/geohazards7010035

with superiority in analyzing and understanding the content's meaning within an accurate semantic context.

Based on that, the proposed framework uses LLMs and DL techniques, with anonymous agents deployed on fog nodes. This wide dispatching enabled the framework to reduce response time compared to the centralized approach while providing greater local context awareness, thereby enhancing the efficiency and reliability of responses to emergency calls.

3. Proposed Methodology

To create comprehensive disaster management, the solution must deal with a disaster before it occurs, during its occurrence, and thereafter. Previous solutions dealt with one side only, like providing a prediction model that only addresses the stage before the disaster. Early detection is not effective in cases where no efficient warning and follow-up are in place. Moreover, the addition of a set of useful rescue services during a crisis would be highly desirable.

This research presents a comprehensive framework to manage disasters, particularly focusing on floods. The proposed framework proposes an enhancement in the accuracy of early prediction models using continuous, real-time monitoring and classification by means of an EL model involving a DL model, a KNN model, and an RF model.

Moreover, the framework collects data at the sensory-level layer from two different resources: an IoT platform for sending and reporting measurement data, in addition to a crowdsourcing mechanism that includes extracting data from a social network (i.e., the X platform) and from a dedicated smart application designed for public use. This double use of resources provides a deeper understanding of disaster impacts based on analyzing users' involvement in reporting observations and comments from the event scene.

In addition, the framework uses a fog layer of gateway devices with computing capability [47] to provide an accurate evaluation of the risk level in each area (covered by a fog node), where each area has a different topological and geographical location. The fog nodes enable the provision of context-based, smarter, and more effective notifications.

To achieve the previous goals, the research proposes two algorithms: the first being a classification algorithm deployed before any disaster occurrence based on IoT-sensed data, while the second algorithm is a TM algorithm for textual data during a disaster. Moreover, the framework provides public users with a smart application with several hot services, in addition to a dashboard that provides real-time statistics and decision-aiding tools.

Based on the data coming from the sensing layer (IoT and crowdsourcing), the fog node applies a first-hand algorithm to assess and determine the perceived current threat level. This process is repeated periodically. When the threat level crosses a threshold, required notifications are sent to the proper stakeholders. Thereafter, the text mining-based method (Method 2) is used to analyze users' tweets and reported observations on the flood level and conditions to show on the control panel of the civil defense team dashboard.

Figure 2 depicts the general view and layers of the proposed framework, while Figure 3 presents a detailed view of the framework. Note that the second algorithm is executed only during a disaster, so alerts are sent only when a threat is detected and classified.

The main components of the proposed framework (see Figure 3):

1. The deployed IoT sensors (water level sensors, temperature and humidity sensors, and wind speed sensors): their mission is to collect data from the actual environment (the city) and send it to the nearest fog node in charge.
2. A platform for data crowdsourcing: a service that enables users to share data with service providers through either a dedicated app (developed specifically) or through social media platforms (e.g., the X tweeting platform).

https://doi.org/10.3390/geohazards7010035

3. Threat-level classifiers: use two automated classification algorithms that identify the threat level based on the rules generated by the EL and TM training models applied to the data coming from the sensing layer.
4. A smart notifier: an intelligent alert model that is responsible for issuing appropriate alerts based on information from the classifiers and using the geolocation context.
5. A multi-model classifier that comprises an EL classifier responsible for identifying classification rules based on historical IoT sensor data, and an LLM-DL classifier that processes and analyzes textual data from the crowdsourcing platform to confirm the EL classifier's decisions.
6. A knowledge database (DB) that is used to store statistical information after the analysis of historical data processed by both classifiers.
7. A geographic information system (GIS) that provides information about the location and topology of the target fog node-controlled area. It focuses on the following criteria: land elevation, whether the land is surrounded by mountains, slope direction and stiffness, the nature of the land's flatness, the availability of water drainage points, and the presence of tunnels.
8. Applications and support services: used for managing alerts, enabling volunteers to participate in data collection, and providing first aid to others.
9. A decision support system (D-Support) that relies on the knowledge base in order to provide useful information for the civil defense teams carrying out disaster management duties.

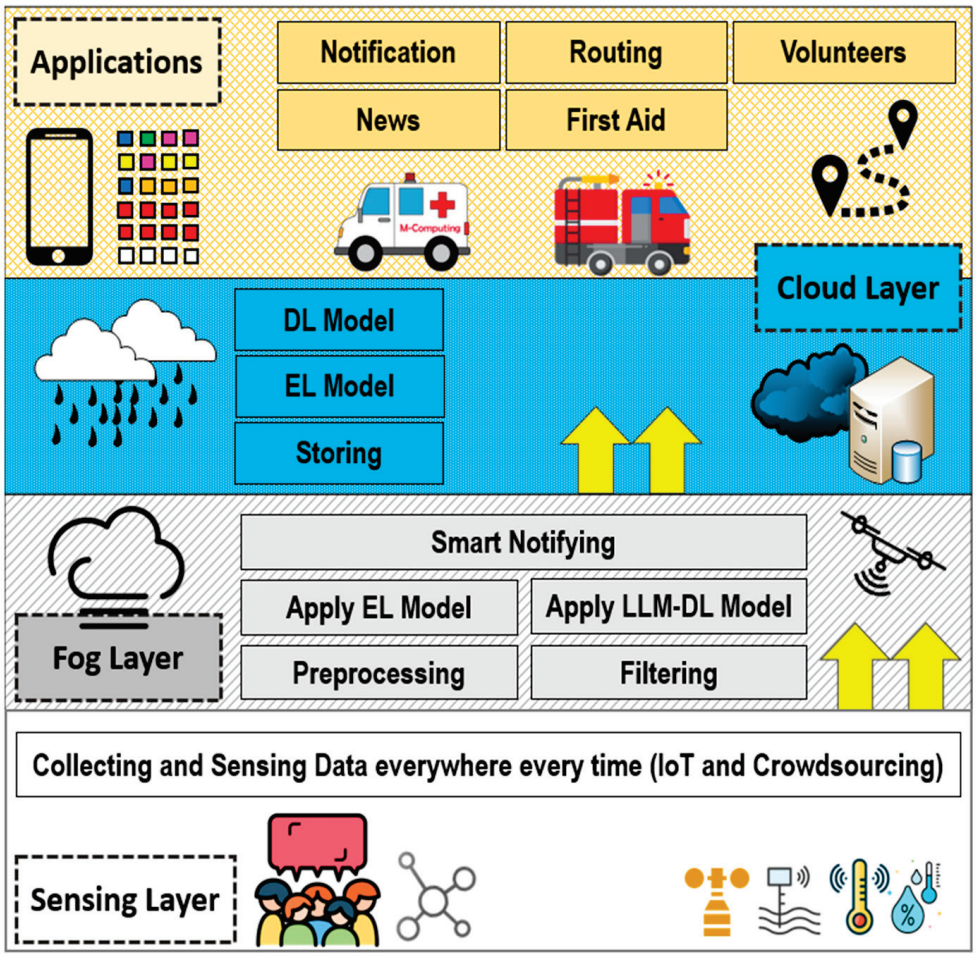


Figure 2. General view of the context-aware classification framework.

https://doi.org/10.3390/geohazards7010035

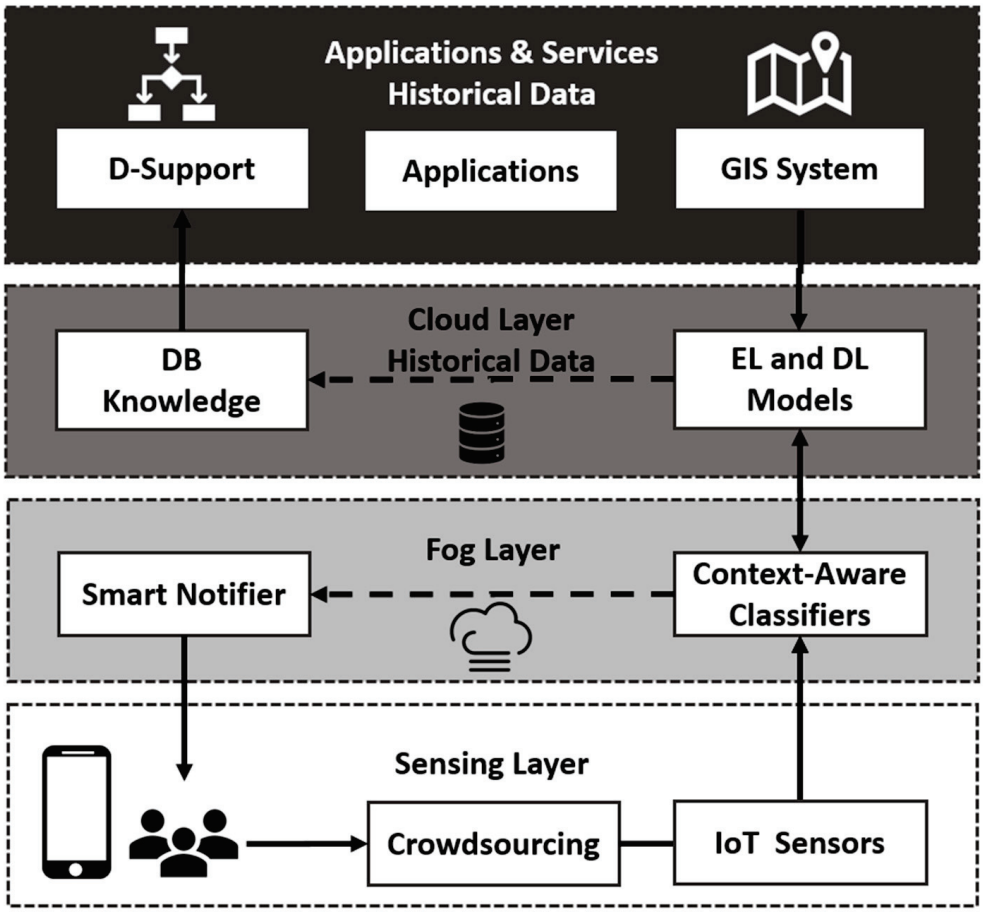


Figure 3. Proposed model and its main components.

3.1. Integration Between Cloud and Fog Computing

The cloud is responsible for processing available collected historical data (with high-performance capabilities). These data were used for training and testing to produce an accurate final model. The final model was then sent to the fog nodes, which continuously test and classify real-time readings based on the model they have received. Periodically, the cloud re-optimized/re-tuned the adopted models using newly collected fresh data that were added to the existing historical dataset. As for the spatial context related to the geographical nature of the area, this was proposed within the research methodology to enhance the proposed framework. However, the testing was done on the whole city area due to the lack of data specific to each region.

Therefore, our conceptual proposal requires that each geographically distinct area (e.g., lowland, highland, floodplain, or areas with or without sewage systems) be managed by a dedicated fog node. Based on this additional information, the risk threshold level in each area can be adjusted according to its specific nature (context-based) in a way that fog nodes will use a context-based threshold level, as opposed to using a unified level in the case of treating the entire city as a single region.

This will improve the classification accuracy in each area based on its unique characteristics. However, this was not implemented in practice, as the available dataset covers the whole city of Jeddah (in Saudi Arabia). In this regard, we started working on a Proof of Concept (PoC) in collaboration with the city's municipality to apply the approach to a specific area within the city of Medina (Saudi Arabia), where stagnant water frequently accumulates in different locations characterized by different topologies.

https://doi.org/10.3390/geohazards7010035

3.2. The EL Classifier (Algorithm 1)

The Internet of Things (IoT) infrastructure comprises many sensors that periodically send their values to enable continuous real-time monitoring of the surrounding environment. In this research, we relied on four types of sensors: water level sensors, temperature sensors, wind speed sensors, and humidity sensors. This selection was made based on consultations with weather experts and after conducting tests on the correlation index between variables in our historical data (over ten years).

The proposed algorithm was trained on historical data for specific sensor readings. The threats were classified in binary format, where 1 indicates the presence of a threat and danger, and 0 indicates no threat. The main idea behind the proposed algorithm is to transform the prediction process into a real-time classification process. In other words, the fog node, responsible for a specific area, performs the classification process for the collected data periodically.

The current situation is classified as either a threat or not. In the case of a threat, the time period is shortened to half, and the process is repeated as needed. If the threat persists, the system speeds up the monitoring process (increasing frequency). In addition, the counter of the threat level of the current fog node-controlled area will be increased by 1. If the counter of the threat level exceeds a certain threshold, appropriate alerts are sent to concerned users, based on their current locations.

Moreover, to enhance the accuracy of the proposed classification algorithm (Algorithm 1), three machine learning models (selected based on empirical experimentations) were adopted as part of the EL model: a Deep Learning model (based on Artificial Neural Networks (ANNs)), a K-nearest Neighbor (KNN) model, and a Random Forest (RF) model. The classification result is based on the majority consensus among the three models. As the ensemble is made up of three models, Algorithm 1 needs at least 2 votes to identify the new sensor read as a threat (majority voting).

Note: Algorithm 1 provides continuous monitoring and alerting indicators for users. Its success is attributed to its ability to leverage fog node coverage, where each node is responsible for a specific area. This allows the fog node to periodically repeat the testing process and make direct adjustments to the threat level counter when appropriate. Furthermore, each area is handled based on its unique geographic nature, enabling the system to send intelligent, customized alerts to each region based on its conditions.

As a matter of fact, the first algorithm (Algorithm 1) operates continuously to monitor the risk level related to rain, heat, wind, and humidity, depending on the contextual nature of each region. If the risk level exceeds a tunable threshold (defined by the weather experts' advice), alerts are sent to the residents (including the volunteer users) of that area. At that point, the system begins monitoring incoming tweets through the application from that specific region to handle and respond to any emergency requests using Method 2.

3.3. The LLM-Based DL Classifier (Method 2)

The importance of crowdsourcing data is that it comes from people at the heart of the event, where human sense and perception could be leveraged to analyze data before sending it to the final classification model. Consequently, the accuracy of the classification during a disaster will improve, and the estimated impact level will be more accurate.

In this endeavor, we propose a smart agent based on an LLM and DL to process textual crowdsourcing data. Regarding the analysis of crowdsourcing data, this work adopts two methods for the textual acquisition of data extracted from:

- Tweets of the X platform, which has proven to be one of the fastest means of news dissemination.

https://doi.org/10.3390/geohazards7010035

- A dedicated proposed mobile smartphone application, which allows volunteer users to send textual information to service providers (SPs) based on their real-world perception of the event.

The received data are preprocessed and analyzed by a real-time, lightweight algorithm used by the LLM, then classified by a DL model to support decision-making for disaster management and civil defense teams. The smart agent calls the LLM API to clean the data and remove stop words. Then, it selects the most important words and matches the highly repeated list of key terms during a disaster. The LLM ensures perfect matches even if users use new synonyms, different languages, or different forms for the same term. After that, the DL classifies each tweet or phrase into two categories (normal, emergency call) and reflects the number of emergency cases on the dashboard.

3.3.1. Phase 1–Creating a List of Key Terms and Training the DL Model

The steps include:

- Collect and label data by an expert (0 normal, 1 threat): the algorithm used a labeled dataset of 1500 tweets [48].
- Apply LLM-API, which will
    - Clean data and remove special characters such as (,:,., etc.) to retain only essential letters. This step, known as "cleaning," is instrumental in preparing text for further analysis.
    - Apply LLM-API to tokenize data and divide the text into distinct words. This step breaks down the text into components to extract meaningful insights. It helps reduce data size and expedite processing.
    - Apply LLM-API to remove stop words. This step is important for eliminating common stop words, such as "the", "is", "to", etc., according to each language.
    - Apply LLM-API to lemmatize each word and find its root without concern about the ISRI or Porter stemmer algorithm that are used in traditional TM. This standardization process ensures the consistency of the analysis data.
    - Build a word cloud to find the most frequent terms as a metric to create a list of key terms. The result of this step will be a list of the most used terms during the flood disaster, which could be validated by a human expert. Then, this expert will validate and refine this list using another list generated by the LLM system without a dataset.
- Create a vector of each tweet in the dataset, which is already classified as normal or a threat.
- Train a DL mode (ANN) on the vectors (30/70 with cross-validation) to create a trained DL model, which will be distributed on fog nodes to retain the context-awareness of location with fast responses.

Hint: the phase 1 steps will be repeated on any new data collected to re-train and enhance the DL model.

3.3.2. Phase 2: Testing Phase (Based on the LLM and DL Model)

The steps include:

- Receive Data: Receiving tweets through the X platform API and users' messages from the dedicated mobile application (volunteers).
- Preprocess data based on LLM-API, match to the list of key terms, and create a vector of the tweet.
- Classify the vector into threat or normal to update the threat level of each spatial context area on the dashboard.

https://doi.org/10.3390/geohazards7010035

The outcome of the above steps is integrated with Algorithm 1 to reflect end results on the dashboard of the civil defense teams, where the results are visualized and tracked geographically for improved decision-making.

3.4. Smart Application and Services

The most important service of the proposed framework is the smart alert service. In addition to its dependence on data from crowdsourcing and IoT, the smart alert service deals with the spatial topological context (to distinguish each area from one another within the city itself). For example, in the case of floods, low-rise and slope-shaped areas that do not contain drainage facilities will be more vulnerable to flooding than high-rise areas. The same is to be said about areas inside tunnels, which also have a high risk of flooding when no proper drainage is present.

The inefficiencies of alert dispatching to users in danger were one of the common shortcomings highlighted in previous work. For example, sending a generic SMS to all city residents, urging caution without specifying actionable steps or hazards to avoid, leads to mistrust in the notification system (especially for those who were not subject to danger at all). This generalization overlooks geographical variations between areas and the relative vulnerability to flood risks. Another defective notification method is posting general warnings on websites that ignore active roles that volunteer users could play by providing accurate data/information about the current flood situation and assisting civil defense teams on the ground.

To address these shortcomings, this work has introduced a smart application designed to deliver specially tailored notifications considering users' location context. Furthermore, beyond its core functionalities, the proposed smart application endeavors to redefine disaster preparedness and response. It proposes providing a spectrum of crucial services to users in times of emergency, covered in the next sub-section. These services are divided into Emergency Alerts and Information Services and User Engagement and Support Services.

A. Emergency Alerts and Information Services:
   - Direct Alert Service: Users receive location-based alerts in real time, ensuring that they get critical information tailored to their current location and to the unique characteristics of the area where each user is. This service is essential for timely and relevant notifications during emergencies.
   - Awareness Service: Users stay informed with periodic articles and notices on what to do and what to avoid during disasters.
   - Status of Areas Service: Users will be able to navigate through the areas that are less dangerous during emergencies and learn how to reach them safely.
   - Road Condition Service: Users will be able to access information about road closures due to disasters and identify available routes.
   - Emergency Numbers: Quick access to essential numbers like civil defense, ambulance, or police.

B. User Engagement and Support Services (Volunteer Users or Defense Teams):
   - Data Sharing: Users can play active roles by sharing real-time data about the location of users to help authorities assess damages and risks accurately. The framework collects and processes these data to verify alert reliability and refine threat level classifications.
   - First Aid: Users can access vital information on how to provide first aid assistance in cases of the delayed arrival of ambulance crews. Users can learn, through text and video resources, how to handle various emergency scenarios, from drowning to bleeding, etc.

https://doi.org/10.3390/geohazards7010035

The enhanced framework includes a control panel (dashboard) that summarizes information about the disaster and facilitates the process of monitoring based on each area's characteristics, relief calls, and the distribution of locations on the map. This, in turn, will support appropriate decision-making and better disaster management by rescue teams. The next section presents the experimental results of the proposed framework.

4. Implementation and Results

This section is organized into five sub-sections that aim to test the main components of the proposed framework to demonstrate its feasibility and effectiveness.

4.1. Testing the Proposed EL Algorithm (Algorithm 1)

The proposed classification algorithm was tested on real data collected over 10 years (from 15 June 2013 to 15 June 2023) from the Saudi Arabian Meteorological Authority for the city of Jeddah. During these years, Jeddah experienced five floods due to rainfall, with the most recent occurring in 2022. The data included daily averages of temperature, humidity, wind speed, sea level elevation, wave height, and rainfall rate. Days that witnessed threats or floods were classified as number 1, while other days were assigned as number 0.

The work was carried out in the Google Colab environment using the Python 3.12 language. After applying the correlation coefficient, the sea level elevation and the sea water level were removed. Preliminary data processing involved replacing some missing values with the median and then applying normalization to all numerical values. Finally, the Synthetic Minority Over-Sampling Technique (SMOTE) was applied to address data imbalance issues, given that the number of threat days was much lower than normal days.

We applied an EL approach, where we selected Deep Learning, Random Forest, and K-nearest Neighbor as our learning ensemble, among several others (all applied to the dataset), as they showed the best accuracy rate, which approached 99% precision. To enhance the accuracy of our proposed algorithm, the majority result among the three models was adopted.

Figure 4 illustrates the comparison results for the previous models, based on the learning curve, and the relevant comparison metrics related to the confusion matrix, such as accuracy, precision, recall, and F1-score, as defined in the following equations:

Accuracy, precision, recall, and F1-score.

$$Accuracy = (TP + TN)/(TP + FP + FN + TN) \quad (1)$$

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}) \quad (2)$$

$$\text{Recall (Sensitivity)} = \text{TP}/(\text{TP} + \text{FN}) \quad (3)$$

$$\text{F1-Score} = 2 * (\text{Recall} * \text{Precision})/(\text{Recall} + \text{Precision}) \quad (4)$$

where TP (True Positive): correctly predicted positive cases, FP (False Positive): incorrectly predicted positive cases (actually negative), FN (False Negative): incorrectly predicted negative cases (actually positive), and TN (True Negative): correctly predicted negative cases. Recall (sensitivity) represents the proportion of actual positive cases (threat) that are correctly identified by the model.

https://doi.org/10.3390/geohazards7010035

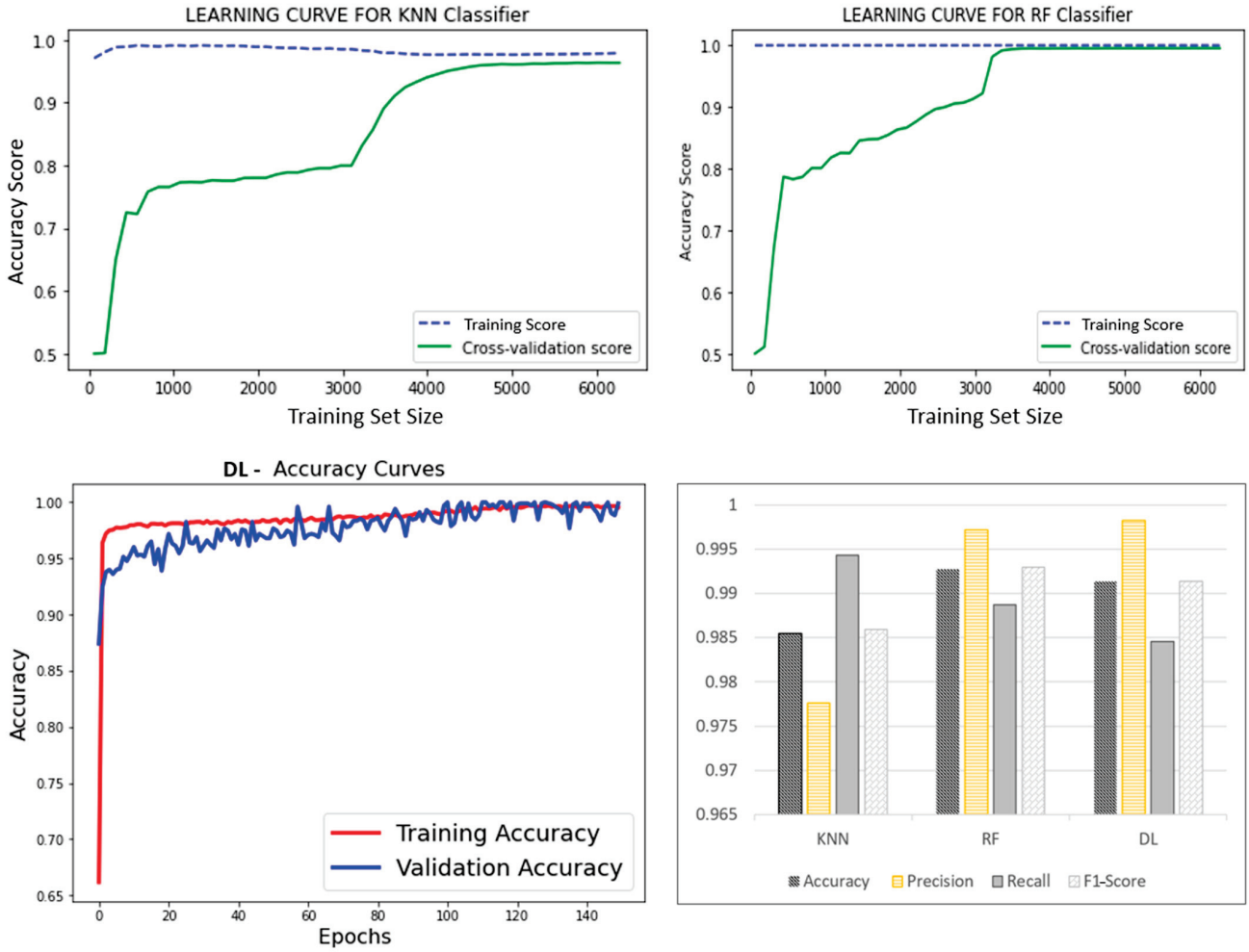


Figure 4. Results of testing the proposed EL-classification algorithm.

4.2. Comparison to Others

This paragraph provides a comparison between our solution and those of some previous studies that deal with flood prediction using ML. Table 3 presents the differences in the dataset size, selected features, selected ML models, and accuracy.

Table 3. Comparison results.

Features of the Dataset					ML Model Accuracy %									Dataset Size	County
					The Selected Models for Our Ensample Learning (KNN, RF, DL)										
Ref.	Month (No.)	Temp	Wind Speed	Rain (mm)	SVM	KNN	RF	LR	DT	DL	Naïve Bayes	DNN	SCV		
[49]	NO	NO	NO	Yes	NO	NO	NO	85.7	57.1	NO	NO	NO	28.57	12,053	Nigeria
[50]	Yes	Yes	NO	Yes	85.6	87.73	NO	NO	NO	NO	85.73	91.18	NO	3120	India
Our	Yes	Yes	Yes	Yes	98.9	99.2	99.6	95.8	99.3	99.1	NO	NO	NO	3654	KSA

In [49], the authors built a model for predicting floods using ML models that depend only on historical rainfall data. A notable limitation was the exclusive dependence on daily rainfall data, neglecting crucial weather parameters such as temperature and wind speed. Moreover, the used dataset is considered old, dating from January 1981 to 31 December 2013. This explains the low accuracy of their results, as evidenced by their recorded metrics: DT = 57.14%, LR 85.7%, and SVM = 28.57%.

In contrast, our dataset was more recent than the datasets used in [49,50]. Also, our study exhibited remarkable advancements in accuracy. For instance, the RF and DL

https://doi.org/10.3390/geohazards7010035

models achieved the highest accuracy, scoring 99.6 percent. This substantial improvement underscores the efficacy of our methodology in mitigating the limitations observed in previous research endeavors.

The authors of [50] incorporated additional weather characteristics into the prediction of flood occurrence, including month, temperature, and rain amount (from 1990 to 2002), yet the accuracy remained modest, with a recorded accuracy of KNN = 85.73 and SVM 85.57. Our study surpassed these by adopting the additional feature of "wind speed".

Moreover, unlike all previous studies, which treated cities as homogeneous entities, our research acknowledges the diverse geographical and climatic characteristics within the city itself. Furthermore, while previous studies predominantly focused on forecasting the amount of rainfall, our research constitutes a paradigm shift toward continuous classification of risk levels for each area in a city.

This strategic divergence allows our model to adapt dynamically to evolving environmental conditions, thereby enhancing their accuracy and practicality. By embracing this innovative approach, our study not only advances the theoretical understanding of flood modeling but also offers practical insights for policymakers and stakeholders tasked with mitigating the impacts of natural disasters.

4.3. Testing the Proposed LLM-Agent Method (Method 2)

This section is dedicated to testing the results of the traditional TM algorithm with an LLM for classifying crowdsourced textual data, whether tweets or comments sent by individuals themselves. A simple dataset of 1500 phrases (tweets for real users in Saudi Arabia between 2014 and 2017) [48] was used and classified as an emergency request or a normal one that does not warrant concern. The algorithm was implemented in Python using the Google Colab platform. The algorithm utilizes the NLTK library for text processing [51]. We used the Confusion Matrix to evaluate the proposed algorithm based on the same criteria. Figure 5 illustrates the results of the experiment. The results show that the accuracy was enhanced from 96% to 99.99% after depending on the LLM.

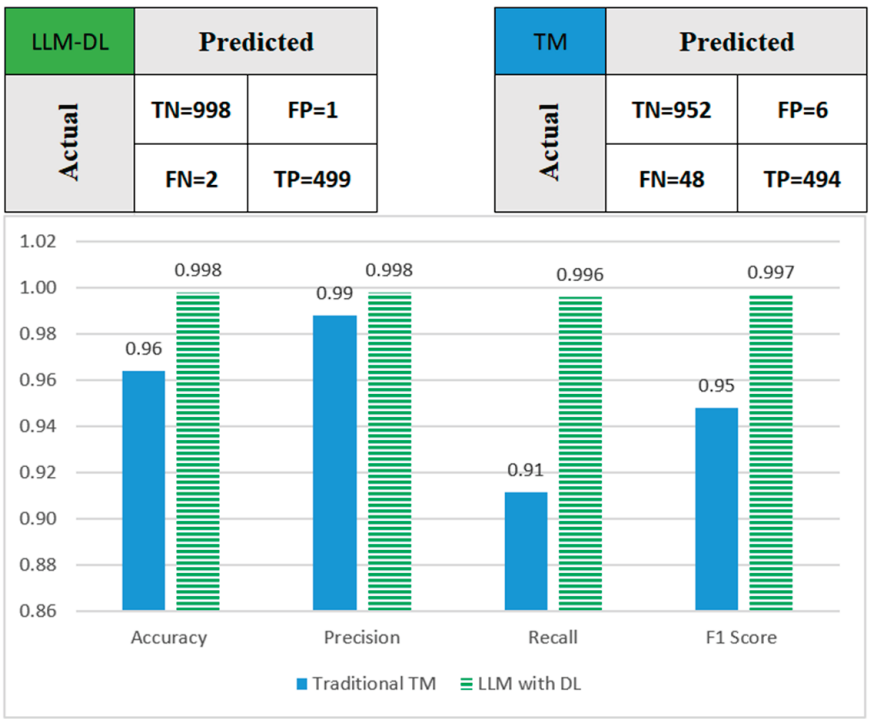


Figure 5. Confusion matrix and comparison between LLM results and traditional TM.

https://doi.org/10.3390/geohazards7010035

4.4. Implementation of the Proposed Application

To ease the evaluation process, a simplified prototype of the proposed application was developed as an Android-based system. The prototype was created using Java for the Android application and PHP for the server-side functionality, utilizing MySQL and Firebase as the database systems. Figures 6 and 7 showcase the primary interfaces (mockups) of the proposed application with dummy data.

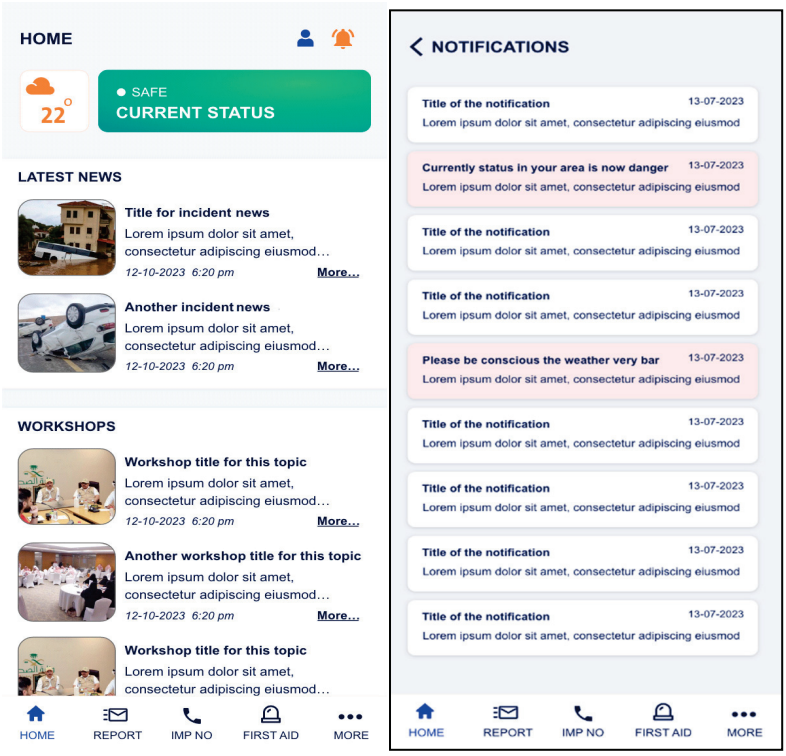


Figure 6. Home screen and notification screen (mockups).

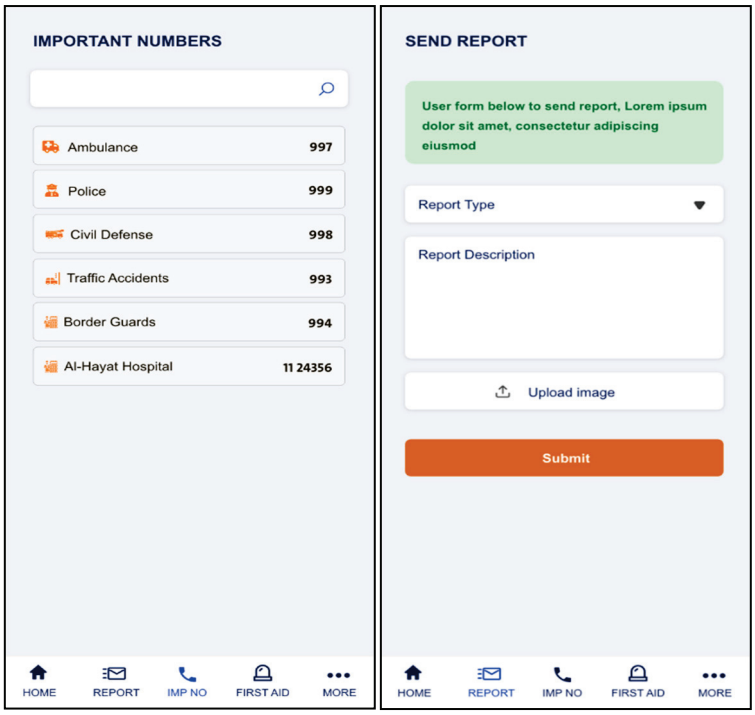


Figure 7. Important number screen and send report screen (volunteer mockups).

https://doi.org/10.3390/geohazards7010035

Figure 6 shows the home screen that displays the main interface of the proposed application, which consists of five distinct sections. The first section includes alerts, the page's title, and the user's profile information. The second section features a weather indicator that provides real-time information on the current level of danger or safety based on the user's location and the current time. The third section is dedicated to delivering crucial news updates. The fourth section provides insights into various workshops, including first aid training, volunteering opportunities, and rescue initiatives. Lastly, the fifth section contains a range of services available within the application. The notification part demonstrates the alerts that are sent to users during times of emergency. The most critical alerts are highlighted with a distinct color scheme to ensure easy identification.

Figure 7 displays the availability of crucial contact information for users to access in urgent situations. This includes direct links to emergency services, like ambulances, civil defense, and traffic management, streamlining swift responses and necessary assistance when required. Moreover, it illustrates the report screen, purposefully designed to streamline data submissions by volunteers and registered users of the application. Volunteers have the option to select the type of disaster, incident, or situation they wish to report. They can provide a detailed description and attach a photo as needed. Additionally, the application automatically includes the location information when forwarding the report to the administration.

4.5. Central Dashboard for Managing Disasters

The dashboard is refreshed in real-time based on the results of the ML and LLM-agent algorithms. Figure 8 provides a simple simulated example of a disaster management dashboard that shows areas according to the degree of danger of the threat and the rates of calls for assistance by people. In addition, the dashboard includes a map showing the distribution of ambulances, rescue teams, drones, and roads that are closed due to flooding. The disaster management team can also send instant alert messages and directions to civil defense teams, volunteers, and people in a specific area by simply selecting the area and entering guidance or a message. The proposed smartphone app is connected to both authorities and end-users.

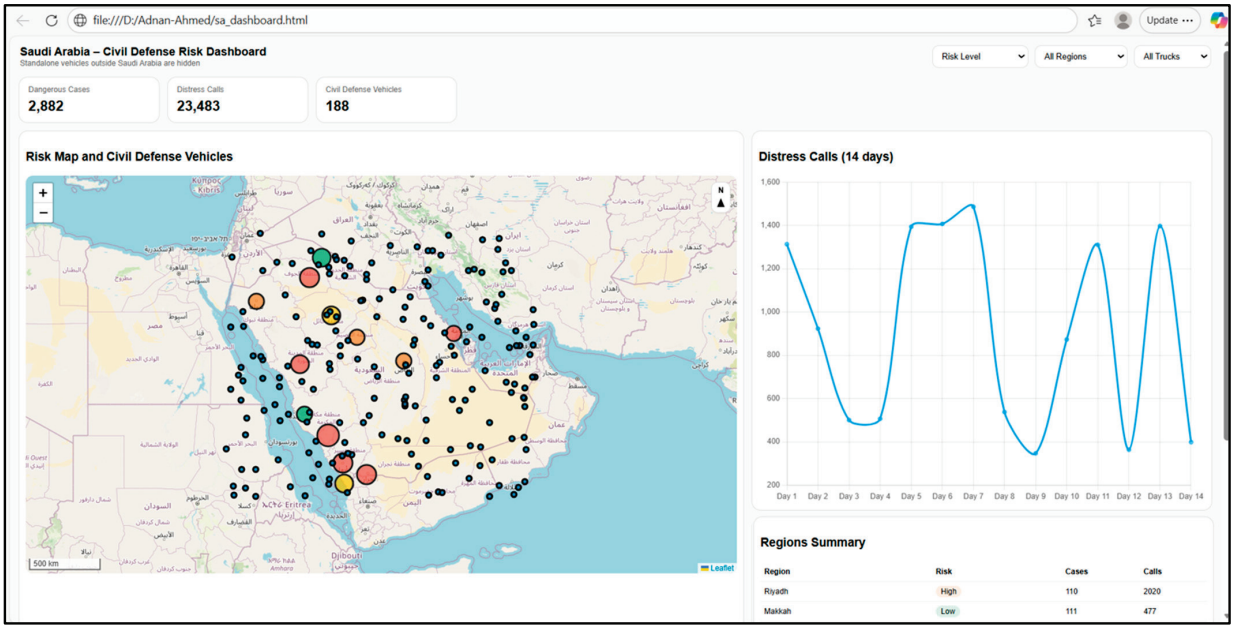


Figure 8. Demo of the proposed dashboard.

https://doi.org/10.3390/geohazards7010035

4.6. Discussion About the Fog Layer Implementation in the City of Madinah

Although this research has not yet deployed the proposed framework in practice, the data used in this study were collected from the Jeddah Meteorological Authority and were used to simulate the operations of the fog layer. Furthermore, for a practical application, we contacted the municipality of Madinah to discuss the application of this research idea in a specific area of the City of Madinah, as a Proof of Concept (PoC), where we agreed to use LORA-based gateways as fog nodes, each one covering an area of 5 km radius. In terms of computational capacity, it is not a matter of much concern since this kind of environmental monitoring does not require heavy computation at fog nodes; thus, an average computer with a reasonable processing capability will suffice. The same could be said about latency; the LoRa technology's affordable latency is acceptable, as it is again an environmental setting where latency is not a stringent constraint [52–54].

4.7. Limitations and Discussion

Similar to any solution, there are some challenges that can form the basis for further refinements and future work. The proposed framework also has a few limitations that this work has tried to provide preliminary solutions for; however, there is still a need for further improvements in future work. These limitations are:

- The collected dataset is geographically and climatologically limited, which may restrict the generalizability of the proposed framework to other regions without retraining. Moreover, flooding patterns, sensor availability, infrastructure resilience, and citizen behavior vary naturally and significantly across regions.
- To relax this challenge, this work uses a DL model as part of the selected models of the EL process. DL provides fast adaptability to changes. Moreover, the proposed framework includes retraining ("on-demand") on new collected data in the cloud to enhance the El model.
- The IoT infrastructure may be vulnerable during severe flooding. The framework mitigates this risk through a multi-layer architecture, sensor redundancy, and integration of crowdsourced data streams as alternative inputs. In addition, we assume the use of waterproof sensors that may be self-powered (e.g., via energy-harvesting platforms [55]) to enable long-term autonomous operations. However, robust communications in disaster cases are an open issue and need new solutions, like satellite connection backup and device-to-device communications, to name a few.
- The integration of Ensemble Learning, Deep Learning, and LLM-based social data analysis introduces considerable computational complexity. While fog computing is proposed to reduce latency, it still faces scalability and resource consumption costs that increase during large-scale disasters and large-scale datasets. However, compared to a centralized system, a distributed system usually offers greater availability and scalability, but at an additional cost that is warranted when dealing with human life.

Moreover, for resource optimization purposes in the municipality's tasks and functions, the IoT and sensor layer could serve other use cases and contribute to different solutions for smart city applications, rather than being confined to the flood/disaster problem alone. Moreover, the fog layer is backed up by a cloud computing layer where complex training and model tuning may happen. As a further enhancement, an aggregated fog layer could be introduced between the cloud and the fog layers to assume aggregation of results, with higher computing power than the fog and lower latency as compared to the cloud.

Furthermore, LoRa system infrastructure is currently affordable (a LoRa tower costs between \$500 and \$1000 US and can cover a region of 1–10 Km^2).

https://doi.org/10.3390/geohazards7010035

The complexity model governing the use of the EL model (involving DL-ANN, RF, and KNN) and LLM at one fog node is expressed as

$$\mathbf{O(EL)} = \mathbf{O(Td + Wl + Nn)} = \mathbf{O(N)} \tag{5}$$

where

n: number of features, which is fixed (five features).
T: number of trees, which is very small at the fog level.
d: level of the tree's depth, which is small with a low number of features.
W: number of weights in each layer, which is fixed after training.
l: number of layers, which is one to three hidden layers with a simple ANN.
N: In general, it has to be the highest value where the KNN recalculates the distance between a new sample and all stored points.

However, this research uses a simple KNN in which the distance is calculated using only two points (the centers of the threat and normal classes), so the computational complexity is not very demanding for a fog computing node.

Regarding LLM complexity, the model is used without training (for semantic analysis), and, therefore, the complexity would be:

$$\mathbf{O(LLM)} = \mathbf{O}\left(\mathbf{T} * \mathbf{w}^2\right) = \mathbf{O(T)} \tag{6}$$

where w is the number of terms in tweets, which is small, i.e., 40–50 words, and T is the number of tweets. Moreover, the tweet analysis is activated only when threat indicators are detected (outside of Algorithm 1 execution). Hence, the total complexity could be expressed as

$$\text{The O(total)} = \text{O(N)} + \text{O(T)} \tag{7}$$

which exhibits a linear trend that is acceptable.

- Another challenge lies in the fact that LLMs with social textual crowdsourcing data analysis are exposed to misinformation, noisy data, sarcasm, multilingual content, or malicious inputs. Consequently, in crisis situations, social media data can be misleading or biased, potentially affecting the accuracy of identifying affected areas.

To mitigate this challenge, our framework is based on a dual use of an IoT-based solution combined with a social data-based analysis.

The main goal of this hybrid solution is to deal with the uncertainty and, sometimes, deceptive nature of social data. Furthermore, the framework provides a dedicated application that enables volunteers to contribute to the monitoring process by sending short messages in addition to using public social data. As a result, this solution guarantees a reduction in data noise and malicious inputs.

Additionally, civil defense officers will be able to examine and identify malicious tweets and bot structures through the dashboard. Reliability is enhanced by averaging tweets (to differentiate between realistic and fake tweets) with a specific confidence threshold. This will reduce the risk of fake tweets, especially since multiple tweets from the same source are ignored.

Finally, it is worth noting that we utilize DL with LLMs to leverage the advantages of semantic language models while avoiding their inherent limitations in the automatic generation process. In addition, fine-tuned prompt engineering with the LLM could be added in the future to further refine LLM responses.

- The framework was tested using collected datasets, but it was not validated through a full real-time pilot deployment during an actual flooding event. This is a major

https://doi.org/10.3390/geohazards7010035

limitation of the study; however, it is common to use available datasets and apply AI solutions to them when no possible real pilot implementation is cost-wise affordable. Moreover, in our current endeavor, we used a real dataset, for which data were collected for 10 years in the same region with multiple repeated flood cases.

- In this manuscript, a smartphone application and a centralized control panel are proposed; the study does not include usability testing or user experience evaluation involving citizens, volunteers, or civil defense teams. The effectiveness of notifications depends heavily on clarity, trust, and user responsiveness, which are not empirically assessed. Our study does not include usability testing or user experience evaluation involving citizens, volunteers, or civil defense teams that adversely impact the effectiveness of notifications. Dealing with this issue will be addressed in a future extension of this work.
- Despite mentioning multiple disaster types in our motivation, the framework is primarily validated for flooding scenarios. Due to limited reliable data and the frequent occurrence of floods in our region in recent years, which are among the most significant natural disasters and threats to the area, this research focuses specifically on floods. However, the proposed framework is general enough to be used for monitoring, detection, and rapid warning of various disasters in the future with different datasets.
- Since the framework is designed for localized processing at fog nodes and the initial testing was conducted using a city-level dataset, the dataset and validation context are characteristic of urban flooding scenarios. Although this research focused primarily on floods, as they have posed the greatest threat in recent years to residents of several cities in Saudi Arabia, largely due to the lack of drainage systems in many areas of the Kingdom, as rain has been scarce. However, the proposed framework is generic in the sense that it could be used to handle multiple types of threats, such as sandstorms, heat waves common to the region, and public health issues.
- The area-specific risk threshold definition is a key aspect of the framework. In this work, the threshold values were determined based on the dataset and local context. In addition, the work relied on the experience of professionals working in civil defense, municipal authorities, and meteorological departments. Their practical knowledge, combined with historical records of past flood events across different areas, will play an important role in setting appropriate risk levels during the real implementation phase.

In particular, lower thresholds will be assigned to areas that have experienced more severe damage during recent floods, as well as to zones with specific topographical characteristics (elevation, rainfall intensity, land use, flow accumulation, slope, and the lag time of the watershed). This will increase sensitivity in high-risk locations.

Finally, to introduce flexibility and adaptability, we proposed the use of a Moving Average Window as a simple and efficient mechanism to allow tuning the dynamic adjustment of risk levels based on new data. However, self-adaptive threshold tuning was not implemented in the current work according to the limitations of the available dataset. Future work will focus on developing more advanced dynamic calibration methods that automatically adjust thresholds based on real-time data patterns and additional inputs.

- A remaining challenge concerns the diversity of regional dialects and colloquial expressions commonly found in emergency-related social media posts. Although the use of LLMs helps address part of this issue, dialect variation (especially in low-resource languages) still requires further research. This issue needs to be addressed further in future work.
- The stability and robustness of the unsupervised LLM component remain important considerations for future work, particularly in ensuring reliable, accountable, and ethically sound automated decision-making in critical emergency management systems. Although this represents an ongoing research challenge, several measures were

https://doi.org/10.3390/geohazards7010035

adopted in the current framework to enhance robustness. These include integrating IoT sensor data with social media inputs to provide cross-validation, employing multiple models within an ensemble to strengthen decision reliability, and also coupling the LLM with Deep Learning to limit the hallucination effects of the LLM while leveraging its semantic understanding capabilities. Further improvements have to focus on strengthening validation mechanisms and governance safeguards.

5. Conclusions

This work introduces a comprehensive framework for managing disasters, with a particular focus on floods, emphasizing early detection. The framework proposes an efficient location and context-aware alert management system. The framework utilizes two sources of data, IoT-systems collected data and crowdsourced data, to manage floods before, during, and after their occurrence. We proposed two algorithms based on Ensemble Learning (EL) and LLMs to classify IoT-related data and textual data into either a genuine threat (for IoT data) or not, and an emergency call or not (for crowdsourced data). Moreover, the fog computing architecture was proposed to facilitate the provision of real-time notifications that are compatible with the spatial context and threat level. Furthermore, we proposed the design of a dedicated smartphone application with various essential services to assist in the collection of textual data (that is, crowdsourced from users and volunteers). The testing of the proposed EL algorithm on a real dataset achieved an impressive accuracy rate of 98%, while the LLM-DL algorithm achieved more than 99% accuracy for textual data.

Finally, the following points may be directions for future work:

- Consider incorporating drone deployment during disaster scenarios for monitoring using image analysis algorithms specifically designed to process data captured by these drones or by surveillance cameras located in specific locations. Moreover, including satellite imagery processing can provide another important source of information in difficult accessibility cases.
- Addressing privacy concerns associated with crowdsourcing and smartphone privacy and security measures will also be a key area of focus.
- One other application of our framework is to apply it to monitor and proactively detect risks of infections caused by prolonged stagnation of water in ponds, leading to diseases such as dengue fever. This is being addressed in collaborative work with the Municipality of Medina to implement the proposed management model, practically on the ground, in a real scenario.

Author Contributions: Conceptualization, A.A.A.S., A.B.M., O.T., N.M.B. and A.B.A.; methodology, A.A.A.S., A.B.M. and N.M.B.; software, A.A.A.S., O.T., N.M.B. and A.B.A.; validation, A.A.A.S., A.B.M., F.H.A., O.T., N.M.B. and A.B.A.; formal analysis, A.A.A.S., A.B.M., F.H.A., O.T., N.M.B. and A.B.A.; investigation, A.A.A.S., A.B.M., O.T., N.M.B. and A.B.A.; resources, A.A.A.S., O.T. and A.B.A.; data curation, A.A.A.S.; writing—original draft preparation, A.A.A.S., A.B.M., O.T., N.M.B. and A.B.A.; writing—review and editing, A.A.A.S., A.B.M., F.H.A., O.T., N.M.B. and A.B.A.; visualization, A.A.A.S., A.B.M., O.T., N.M.B. and A.B.A.; supervision, A.A.A.S.; project administration, A.A.A.S.; funding acquisition, O.T. All authors have read and agreed to the published version of the manuscript.

Funding: This paper is derived from a research grant funded by the Research, Development and Innovation Authority (RDIA)–Kingdom of Saudi Arabia–with grant number (13465-upm-2023-upm-R-3-1-EF-) as well as the Osamah AlSayyed research center fund award.

Data Availability Statement: The datasets generated and analyzed during the current study are available in the [GitHub] repository, [https://github.com/adnanmnm/Flood-Dataset], available online as of 28 February 2026.

Conflicts of Interest: The authors declare no conflicts of interest.

https://doi.org/10.3390/geohazards7010035

References

1. Tuzyak, Y.M.; Tuzyak, O. Risk management and advanced systems for observing, monitoring and forecasting natural disasters and events. In *Third EAGE Workshop on Assessment of Landslide Hazards and Impact on Communities*; European Association of Geoscientists & Engineers: Bunnik/Houten, The Netherlands, 2021; Volume 2021, pp. 1–5.
2. Tomar, P.; Singh, S.K.; Kanga, S.; Meraj, G.; Kranjčić, N.; Đurin, B.; Pattanaik, A. GIS-based urban flood risk assessment and management—A case study of Delhi National Capital Territory (NCT), India. *Sustainability* **2021**, *13*, 12850. [CrossRef]
3. Luu, C.; Tran, H.X.; Pham, B.T.; Al-Ansari, N.; Tran, T.Q.; Duong, N.Q.; Dao, N.H.; Nguyen, L.P.; Nguyen, H.D.; Thu Ta, H.; et al. Framework of spatial flood risk assessment for a case study in Quang Binh province, Vietnam. *Sustainability* **2020**, *12*, 3058. [CrossRef]
4. Christian Aid. Counting the Cost 2020: A Year of Climate Breakdown. December 2020. Available online: https://www.christianaid.org.uk/sites/default/files/2022-12/counting-the-cost-2022.pdf (accessed on 28 February 2026).
5. Xu, H.; Wei, W.; Qi, Y.; Qi, S. Blockchain-Based Crowdsourcing Makes Training Dataset of Machine Learning No Longer Be in Short Supply. *Wirel. Commun. Mob. Comput.* **2022**, *2022*, 7033626. [CrossRef]
6. Le, T.; Ngoc, T. Floods and household welfare: Evidence from Southeast Asia. *Econ. Disasters Clim. Chang.* **2020**, *4*, 145–170. [CrossRef]
7. Plazas, J.E.; Bimonte, S.; Schneider, M.; De Vaulx, C.; Battistoni, P.; Sebillo, M.; Corrales, J.C. Sense, Transform & Send for the Internet of Things (STS4IoT): UML profile for data-centric IoT applications. *Data Knowl. Eng.* **2022**, *139*, 101971.
8. Aljohani, F.H.; Alkhodre, A.B.; Sen, A.A.A.; Ramazan, M.S.; Alzahrani, B.; Siddiqui, M.S. Flood Prediction using Hydrologic and ML-based Modeling: A Systematic Review. *Int. J. Adv. Comput. Sci. Appl.* **2023**, *14*, 538–551. [CrossRef]
9. Hu, H.; Yang, H.; Wen, J.; Zhang, M.; Wu, Y. An Integrated Model of Pluvial Flood Risk and Adaptation Measure Evaluation in Shanghai City. *Water* **2023**, *15*, 602. [CrossRef]
10. Gulati, K.; Boddu, R.S.K.; Kapila, D.; Bangare, S.L.; Chandnani, N.; Saravanan, G. A review paper on wireless sensor network techniques in Internet of Things (IoT). *Mater. Today Proc.* **2022**, *51*, 161–165. [CrossRef]
11. Tan, W.; Sidhu, M.S. Review of RFID and IoT integration in supply chain management. *Oper. Res. Perspect.* **2022**, *9*, 100229. [CrossRef]
12. Tayan, O. Context-Aware Framework for Enhanced Smart Urban Pollution Monitoring and Control. In *2022 International Conference on Emerging Trends in Computing and Engineering Applications (ETCEA)*; IEEE: New York, NY, USA, 2022; pp. 1–5.
13. Esparza, M.; Farahmand, H.; Brody, S.; Mostafavi, A. Examining data imbalance in crowdsourced reports for improving flash flood situational awareness. *Int. J. Disaster Risk Reduct.* **2023**, *95*, 103825. [CrossRef]
14. Tashtoush, Y.; Alrababah, B.; Darwish, O.; Maabreh, M.; Alsaedi, N. A deep learning framework for detection of COVID-19 fake news on social media platforms. *Data* **2022**, *7*, 65. [CrossRef]
15. Jaradat, S.; Nayak, R.; Paz, A.; Ashqar, H.I.; Elhenawy, M. Multitask learning for crash analysis: A Fine-Tuned LLM framework using Twitter data. *Smart Cities* **2024**, *7*, 2422–2465. [CrossRef]
16. Heaton, D.; Nichele, E.; Clos, J.; Fischer, J.E. "ChatGPT says no": Agency, trust, and blame in Twitter discourses after the launch of ChatGPT. *AI Ethics* **2025**, *5*, 653–675.
17. Donkers, T.; Ziegler, J. Understanding Online Polarization Through Human-Agent Interaction in a Synthetic LLM-Based Social Network. In Proceedings of the International AAAI Conference on Web and Social Media, Copenhagen, Denmark, 23–26 June 2025; Volume 19, pp. 457–478.
18. Ghali, M.K.; Farrag, A.; Lam, S.; Won, D. BEYONDWORDS is All You Need: Agentic Generative AI based Social Media Themes Extractor. *arXiv* **2025**, arXiv:2503.01880.
19. Kumar, V.; Sharma, K.V.; Caloiero, T.; Mehta, D.J.; Singh, K. Comprehensive Overview of Flood Modeling Approaches: A Review of Recent Advances. *Hydrology* **2023**, *10*, 141. [CrossRef]
20. Hosseiny, H.; Nazari, F.; Smith, V.; Nataraj, C. A framework for modeling flood depth using a hybrid of hydraulics and machine learning. *Sci. Rep.* **2020**, *10*, 8222. [CrossRef]
21. Cai, S.; Fan, J.; Yang, W. Flooding risk assessment and analysis based on gis and the tfn-ahp method: A case study of chongqing, china. *Atmosphere* **2021**, *12*, 623. [CrossRef]
22. Adeel, A.; Gogate, M.; Farooq, S.; Ieracitano, C.; Dashtipour, K.; Larijani, H.; Hussain, A. A survey on the role of wireless sensor networks and IoT in disaster management. *Geol. Disaster Monit. Based Sens. Netw.* **2018**, *3*, 57–66.
23. Baky, M.A.A.; Islam, M.; Paul, S. Flood hazard, vulnerability and risk assessment for different land use classes using a flow model. *Earth Syst. Environ.* **2020**, *4*, 225–244. [CrossRef]
24. Abdelkarim, A.; Gaber, A.F. Flood risk assessment of the wadi nu'man basin, mecca, Saudi Arabia (during the period, 1988–2019) based on the integration of geomatics and hydraulic modeling: A case study. *Water* **2019**, *11*, 1887. [CrossRef]
25. Ghile, H.; Shirakawa, H.; Tanikawa, H. Application of GIS and machine learning to predict flood areas in Nigeria. *Sustainability* **2022**, *14*, 5039. [CrossRef]

https://doi.org/10.3390/geohazards7010035

26. Daoudi, M.; Niang, A.J. Flood risk and vulnerability of jeddah city, saudi arabia. In *Recent Advances in Flood Risk Management*; IntechOpen: London, UK, 2019; pp. 634–654.
27. Ullah, K.; Zhang, J. Gis-based flood hazard mapping using relative frequency ratio method: A case study of panjkora river basin, eastern hindu kush, Pakistan. *PLoS ONE* **2020**, *15*, e0229153. [CrossRef]
28. Aljohani, H.; Sen, A.A.A.; Ramazan, M.S.; Alzahrani, B.; Bahbouh, N.M. A Smart Framework for Managing Natu-ral Disasters Based on the IoT and ML. *Appl. Sci.* **2023**, *13*, 3888. [CrossRef]
29. Ke, Q.; Tian, X.; Bricker, J.; Tian, Z.; Guan, G.; Cai, H.; Huang, X.; Yang, H.; Liu, J. Urban pluvial flooding prediction by machine learning approaches—A case study of Shenzhen city, China. *Adv. Water Resour.* **2020**, *145*, 103719. [CrossRef]
30. Arabameri, A.; Saha, S.; Mukherjee, K.; Blaschke, T.; Chen, W.; Ngo, P.T.T.; Band, S.S. Modeling spatial flood using novel ensemble artificial intelligence approaches in northern Iran. *Remote Sens.* **2020**, *12*, 3423. [CrossRef]
31. Karimiziarani, M.; Jafarzadegan, K.; Abbaszadeh, P.; Shao, W.; Moradkhani, H. Hazard risk awareness and disaster management: Extracting the information content of twitter data. *Sustain. Cities Soc.* **2022**, *77*, 103577. [CrossRef]
32. Bryan-Smith, L.; Godsall, J.; George, F.; Egode, K.; Dethlefs, N.; Parsons, D. Real-time social media sentiment analysis for rapid impact assessment of floods. *Comput. Geosci.* **2023**, *178*, 105405. [CrossRef]
33. Yuan, F.; Fan, C.; Farahmand, H.; Coleman, N.; Esmalian, A.; Lee, C.-C.; Patrascu, F.I.; Zhang, C.; Dong, S.; Mostafavi, A. Smart flood resilience: Har-nessing community-scale big data for predictive flood risk monitoring, rapid impact assessment, and situational awareness. *Environ. Res. Infrastruct. Sustain.* **2022**, *2*, 025006. [CrossRef]
34. Singh, J.P.; Dwivedi, Y.K.; Rana, N.P.; Kumar, A.; Kapoor, K.K. Event classification and location prediction from tweets during disasters. *Ann. Oper. Res.* **2019**, *283*, 737–757. [CrossRef]
35. Jongman, B.; Wagemaker, J.; Romero, B.R.; de Perez, E.C. Early flood detection for rapid humanitarian response: Harnessing near real-time satellite and Twitter signals. *ISPRS Int. J. Geo-Inf.* **2015**, *4*, 2246–2266. [CrossRef]
36. Kanth, K.; Chitra, P.; Sowmya, G.G. Deep learning-based assessment of flood severity using social media streams. *Stoch. Environ. Res. Risk Assess.* **2022**, *36*, 473–493. [CrossRef]
37. Liu, X.; Kar, B.; Ishino, F.A.M.; Zhang, C.; Williams, F. Assessing the reliability of relevant tweets and validation using manual and automatic approaches for flood risk communication. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 532. [CrossRef]
38. Cicek, D.; Kantarci, B. Use of Mobile Crowdsensing in Disaster Management: A Systematic Review, Challenges, and Open Issues. *Sensors* **2023**, *23*, 1699. [CrossRef] [PubMed]
39. Tripathy, S.S.; Chaudhuri, S.; Murtugudde, R.; Mhatre, V.; Parmar, D.; Pinto, M.; Zope, P.E.; Dixit, V.; Karmakar, S.; Ghosh, S. Analysis of Mumbai Floods in recent Years with Crowdsourced Data. *arXiv* **2023**, arXiv:2306.09770. [CrossRef]
40. Podhoranyi, M. A comprehensive social media data processing and analytics architecture by using big data platforms: A case study of twitter flood-risk messages. *Earth Sci. Inform.* **2021**, *14*, 913–929. [CrossRef] [PubMed]
41. Alabbas, W. Classification of colloquial Arabic tweets in real-time to detect high-risk floods. In Proceedings of the 2017 International Conference On Social Media, Wearable and Web Analytics (Social Media), London, UK, 19–20 June 2017.
42. Kankanamge, N.; Yigitcanlar, T.; Goonetilleke, A.; Kamruzzaman, M. Determining disaster severity through social media analysis: Testing the methodology with Southeast Queensland Flood tweets. *Int. J. Disaster Risk Reduct.* **2020**, *42*, 101360. [CrossRef]
43. Kumar, A.; Singh, J.P.; Rana, N.P.; Dwivedi, Y.K. Multi-Channel Convolutional Neural Network for the Identification of Eyewitness Tweets of Disaster. *Inf. Syst. Front.* **2022**, *25*, 1589–1604. [CrossRef]
44. Young, J.; Arthur, R.; Spruce, M.; Williams, H.T. Social sensing of flood impacts in India: A case study of Kerala 2018. *Int. J. Disaster Risk Reduct.* **2022**, *74*, 102908. [CrossRef]
45. Caballero, A.; Centeno, R.; Rodrigo, Á. LLM-Based Multi-Agent Models for Multiclass Classification of Strategic Narratives. In Proceedings of the IberLEF 2024, Valladolid, Spain, 24 September 2024.
46. Linardos, V.; Drakaki, M.; Tzionas, P. Utilizing LLMs and ML Algorithms in Disaster-Related Social Media Content. *GeoHazards* **2025**, *6*, 33. [CrossRef]
47. Sen, A.A.; Yamin, M. Advantages of using fog in IoT applications. *Int. J. Inf. Technol.* **2021**, *13*, 829–837. [CrossRef]
48. Hamoui, B.; Mars, M.; Almotairi, K. FloDusTA: Saudi tweets dataset for flood, dust storm, and traffic accident events. In Proceedings of the Twelfth Language Resources and Evaluation Conference, Marseille, France, 11–16 May 2020; pp. 1391–1396. Available online: https://aclanthology.org/2020.lrec-1.174/ (accessed on 28 February 2026).
49. Lawal, Z.K.; Yassin, H.; Zakari, R.Y. Flood prediction using machine learning models: A case study of kebbi state Nigeria. In *2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*; IEEE: New York, NY, USA, 2021; pp. 1–6.
50. Sankaranarayanan, S.; Prabhakar, M.; Satish, S.; Jain, P.; Ramprasad, A.; Krishnan, A. Flood prediction based on weather parameters using deep learning. *J. Water Clim. Change* **2020**, *11*, 1766–1783.
51. Yeasmin, N.; Mahbub, N.I.; Baowaly, M.K.; Singh, B.C.; Alom, Z.; Aung, Z.; Azim, M.A. Analysis and prediction of user sentiment on COVID-19 pandemic using tweets. *Big Data Cogn. Comput.* **2022**, *6*, 65. [CrossRef]
52. Lv, L.; Wu, Z.; Zhang, J.; Zhang, L.; Tan, Z.; Tian, Z. A VMD and LSTM based hybrid model of load forecasting for power grid security. *IEEE Trans. Ind. Inform.* **2021**, *18*, 6474–6482. [CrossRef]

https://doi.org/10.3390/geohazards7010035

53. Lv, L.; Wu, Z.; Zhang, L.; Gupta, B.B.; Tian, Z. An edge-AI based forecasting approach for improving smart microgrid efficiency. *IEEE Trans. Ind. Inform.* **2022**, *18*, 7946–7954. [CrossRef]
54. Gia, T.N.; Queralta, J.P.; Westerlund, T. Exploiting LoRa, edge, and fog computing for traffic monitoring in smart cities. In *LPWAN Technologies for IoT and M2M Applications*; Academic Press: Cambridge, MA, USA, 2020; pp. 347–371.
55. Crescini, D.; Touati, F.; Crescini, P.; Legena, C.; Galli, A.; Mnaouer, A.B. Multi-Parametric Environmental Diagnostics and Monitoring Sensor Node. U.S. Patent No. 10,429,367, 1 October 2019.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

https://doi.org/10.3390/geohazards7010035

big data and cognitive computing

MDPI

Article

An Intelligent Evaluation Method for Slope Stability Based on a Database Integrating Real Cases and Numerical Simulations

Junyi Jiang [1], Dong Li [1], Qingyi Yang [1], Zhenhua Zhang [2], Lei Wang [2], Wenru Zhao [2] and Mingliang Chen [2,3,*]

[1] Shandong Electric Power Engineering Consulting Institute Corp., Ltd., Jinan 250013, China
[2] School of Civil Engineering, Hefei University of Technology, Hefei 230009, China
[3] State Key Laboratory of Hydraulics and Mountain River Engineering, Sichuan University, Chengdu 610065, China
* Correspondence: 2023800025@hfut.edu.cn

Abstract

Slope instability can cause severe disasters, making stability prediction essential. Machine learning has become a key tool for this purpose, as it avoids complex mechanical calculations and efficiently handles high-dimensional data. Currently, the data used in machine learning primarily originate from real-world cases. However, such cases are inherently limited in quantity and often fail to comprehensively represent all potential slope conditions. To address these limitations, this study proposes a method for constructing numerical simulation databases. Based on this, we develop a model establishment method for rapid evaluation of slope stability integrating numerical simulation with engineering cases. This study uses six characteristic parameters to assess slope stability, including unit weight γ, cohesion c, internal friction angle φ, slope angle α, slope height H, and pore pressure ratio r_u. Through extensive literature mining, we established a database of 684 engineering cases. Based on statistical analysis of input parameters, a numerical simulation scheme was designed. Batch calculations were performed using MATLAB to determine simulation results. The engineering case database was then partitioned into training and testing sets for model development and validation. Subsequently, the numerical simulation database was incorporated into the training set for retesting. Results demonstrate that when considering all predictive indicators, the prediction accuracy of the GRNN-based model improved from 85% to 88.3%, while the PNN-based model showed an increase from 69% to 88.3%. This study offers new insights for optimizing numerical simulation design and enhancing machine learning performance in slope stability prediction.

Keywords: machine learning; neural networks; slope stability prediction; model training

1. Introduction

Large-scale anthropogenic slopes are commonly formed through various engineering activities such as open-pit mining operations, water conservancy projects, land reclamation initiatives and infrastructure construction. Slopes act as complex geoengineering systems and their stability presents prominent complexity and uncertainty under the influence of multiple interrelated factors. Slope failure not only brings about severe engineering safety risks but also has the potential to induce major accidents, which in turn lead to immeasurable socioeconomic losses and adverse environmental consequences. Prominent instances of such disasters include the Bijie landslide in Guizhou on 3 January 2022 [1], the catastrophic Shuicheng landslide in Guizhou on 23 July 2019 [2], the open-pit mine

collapse at Xinjing Coal Mining Co., Ltd. in Alxa Left Banner, Inner Mongolia on 2 February 2023 [3] and the Sierra Leone landslide on 14 August 2017 [4]. All these incidents caused heavy casualties and significant economic losses. It is thus of crucial importance to apply modern technologies effectively for the prediction and control of slope stability. With the continuous progress of computer technology, machine learning methods have been gradually introduced into the field of slope stability prediction, which allows for more intelligent and precise analytical work in this regard.

Current intelligent prediction methods for slope stability research have been explored through a variety of machine learning-based studies. Samui et al. (2011) adopted a least squares support vector machine model to conduct analysis and research on slope stability, and further compared its performance with that of artificial neural networks [5]. Their research outcomes provide the engineering field with a reliable analytical method for slope stability assessment. Gordan et al. (2016) developed a particle swarm optimization-based artificial neural network model, which exhibits superior predictive performance relative to traditional artificial neural networks [6]. Gao et al. (2020) proposed an artificial neural network model optimized by the imperialist competitive algorithm, and this model achieves higher prediction accuracy for the stability of cohesive soil slopes compared with conventional artificial neural network methods [7]. Zhou et al. (2019) applied the gradient boosting machine method to slope stability prediction, and the method shows stronger adaptability to complex data processing and higher prediction accuracy than traditional approaches [8]. Bui et al. (2020) developed an M5Rules-GA model based on the decision tree algorithm, and comprehensive evaluation metrics have verified the superior predictive performance of this model [9]. Deng et al. (2020) improved the classical extreme learning machine and proposed a forgetting factor-based online sequential extreme learning machine model [10]. Comparative studies have confirmed that this model outperforms traditional extreme learning machine methods in predictive performance. Moayedi et al. (2021) integrated Harris Hawks Optimization with artificial neural networks to develop a novel prediction model for slope stability [11]. This hybrid HHO-ANN method significantly improves the prediction accuracy and reliability of traditional Multilayer Perceptron techniques in the stability analysis of soil slopes. In addition, numerous researchers have also adopted hybrid algorithm models for slope stability prediction. Sun et al. (2022) developed a stacked ensemble learning model optimized by Bayesian methods and applied it to the analysis of 210 landslide cases [12]. This method avoids complex calculation processes and provides valuable references for the construction of intelligent slope identification platforms. Zhang et al. (2022) proposed an ensemble learning method combining Random Forest and Extreme Gradient Boosting for slope stability prediction [13]. This method offers an effective solution for the accurate characterization of slope states and can be further extended to the stability evaluation of landslide-prone areas.

Existing research reveals that the machine learning models currently applied in slope stability analysis mostly rely on limited engineering case data, which is subject to such constraints as insufficient sample size and incomplete coverage of slope types and working conditions [14,15]. Although some scholars have attempted to expand the sample database by means of numerical simulation methods, the parameter settings in these simulation processes are often only based on the statistical characteristics of existing engineering data, leading to questionable rationality of the generated simulated data. For instance, Ma et al. (2025) conducted numerical simulations using FLAC3D and constructed a database of 96 samples, but the simulation parameters were randomly assigned within broad engineering ranges without referencing the statistical characteristics of real engineering cases [16]. Li et al. (2023) batch-generated 880 homogeneous slope samples via FLAC3D and calculated the factor of safety (FoS) for each; however, samples with unobtainable FoS were

https://doi.org/10.3390/bdcc10030087

directly classified as unstable without verifying the physical rationality of their parameter combinations [17]. Khan et al. (2026) generated 249 simulated samples through limit equilibrium analyses in GeoStudio, but the parameter design was confined to the specific geometric configuration of the Meizhou landslide, resulting in poor generalizability to other slope types [18]. Hsiao et al. (2022) produced slope samples using the random finite element method, yet their parameter selection only considered partial statistical features (e.g., mean and range) of engineering cases, ignoring inter-parameter dependency patterns and distribution density characteristics [19]. By integrating comprehensive statistical analysis, dual-constraint sampling, and physical validity verification, the numerical simulation database constructed in this study overcomes the limitations of previous datasets. It provides a more robust foundation for training machine learning models with strong generalization capabilities, enabling their application in real-world engineering scenarios. A method for constructing a numerical simulation database is first developed to effectively expand the training dataset and make up for the deficiency of measured engineering data. On this basis, a rapid modeling method for slope stability evaluation is further proposed, which organically integrates numerical simulation data with actual engineering case data for model construction and verification.

2. Data

2.1. Data Collection

To build a comprehensive and accurate database of actual slope cases, this study carried out extensive literature mining and removed duplicate records, finally compiling 684 datasets to form the engineering case database with 352 stable slope cases and 332 failed slope cases included. The data were mainly derived from the references cited below [20,21], and the complete slope case database is provided in the Supplementary Materials.

Slope sliding typically occurs in the vicinity of the potential slip surface. Under the combined influence of engineering factors and natural conditions, the shear stress acting on the potential slip surface exceeds its shear strength, which leads to local slope instability as shown in Figure 1. A synthesis of empirical evidence from engineering practice and findings from theoretical research shows that slope stability is mainly governed by three key factors, namely the geomechanical properties of potential failure surfaces, the basic geometric configurations of slopes and exogenous triggering mechanisms [22,23]. The predictive parameters selected for slope stability assessment were determined based on their statistical independence and practical measurability in field investigations, with unit weight γ, cohesion c, internal friction angle φ, slope angle α, slope height H and pore pressure ratio r_u included as the core parameters.

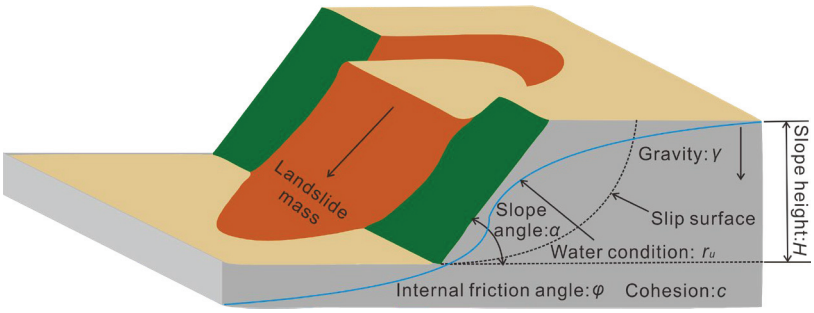


Figure 1. The 3D schematic diagram of slope failure.

Unit weight γ is the weight per unit volume of the geotechnical mass, which directly determines the gravitational potential energy of the slope and the size of the sliding force. According to the limit equilibrium theory of slope stability analysis, the sliding force is

https://doi.org/10.3390/bdcc10030087

positively correlated with the unit weight of the geotechnical mass, and the larger the unit weight is, the larger the shear force along the potential sliding surface of the slope is, and the worse the stability is [24]. In the Mohr–Coulomb strength criterion ($\tau = c + \sigma_v \cdot \tan \varphi$), the unit gravity indirectly controls the shear strength by influencing the normal stress, and also serves as the basis for constructing the self-gravitating stress field of the slope mass ($\sigma_v = \gamma \cdot H$). In the sliding analysis of natural slopes, the self-gravitating stress field usually plays a dominant role in the stress distribution on the potential sliding surface [25].

Cohesion c is the inherent shear strength formed between geotechnical particles through cementation and molecular gravity, which is the ability of the geotechnical mass to resist shear damage in the stress-free state and exists independently of the normal stress. As one of the core parameters of the Mohr–Coulomb strength criterion, cohesion directly contributes to the shear strength τ [23]. The greater the cohesion, the greater the intrinsic ability of the slope to resist sliding [26]. The internal friction angle φ reflects the friction effect and occlusion between the geotechnical particles, and represents the slip resistance of the geotechnical mass under normal stress [23]. In the Mohr–Coulomb criterion, the internal friction angle φ determines the enhancement effect of normal stress on shear strength through the $\sigma \cdot \tan \varphi$ term [27]. These two parameters are the core indicators of the shear strength of a geotechnical mass, which is the decisive factor in controlling the stability [28].

The slope angle α is a key geometric parameter that determines the degree of slope inclination and the distribution of shear stresses on the potential sliding surface. According to the limit equilibrium theory, the tangential component of the slope self-weight stress ($\gamma \cdot H \cdot \sin \alpha \cdot \cos \alpha$) increases with increasing α, while the normal stress ($\gamma \cdot H \cdot \cos^2 \alpha$) decreases. The ratio of shear force relative to shear strength increases sharply as the slope angle increases [28,29]. Slope angle is the most intuitive parameter that affects slope stability.

The slope height H determines the scale of the self-gravitating stress field of the slope, and the maximum self-gravitating stress can be calculated according to $\sigma_v = \gamma \cdot H$. In slopes, an increase in slope height leads to a general increase in the self-gravitational stresses on the potential sliding surface [22,28]. Particularly critical is that this causes the sliding force to increase faster than the slip resistance, since cohesion provides a contribution to slip resistance proportional to H, whereas the sliding force is approximately proportional to H^2 [24].

The weakening effect of pore water pressure on the strength of a geotechnical mass can be quantified by the pore pressure ratio ($r_u = u/\sigma_v$). The essence of this is a reduction in the effective normal stress ($\sigma' = \sigma - u$) that controls the shear strength, which directly leads to a decrease in the Mohr–Coulomb strength of the material ($\tau = c + \sigma' \tan \varphi$). Therefore, r_u is a decisive parameter for evaluating the stability of slopes under seepage conditions [12,26,30]. Scenarios such as rainfall and groundwater rise can lead to an increase in pore water pressure, which reduces the effective stress and drastically decreases the shear strength, making it highly susceptible to landslides.

2.2. Data Analysis

The parameters for slope stability analysis include unit weight, cohesion, internal friction angle, slope angle, slope height, and pore pressure ratio. The violin plots of the engineering case database clearly reveal the distribution characteristics of each parameter. As shown in Figure 2, the violin plot is an advanced visualization tool that combines box plots with kernel density estimation.

https://doi.org/10.3390/bdcc10030087

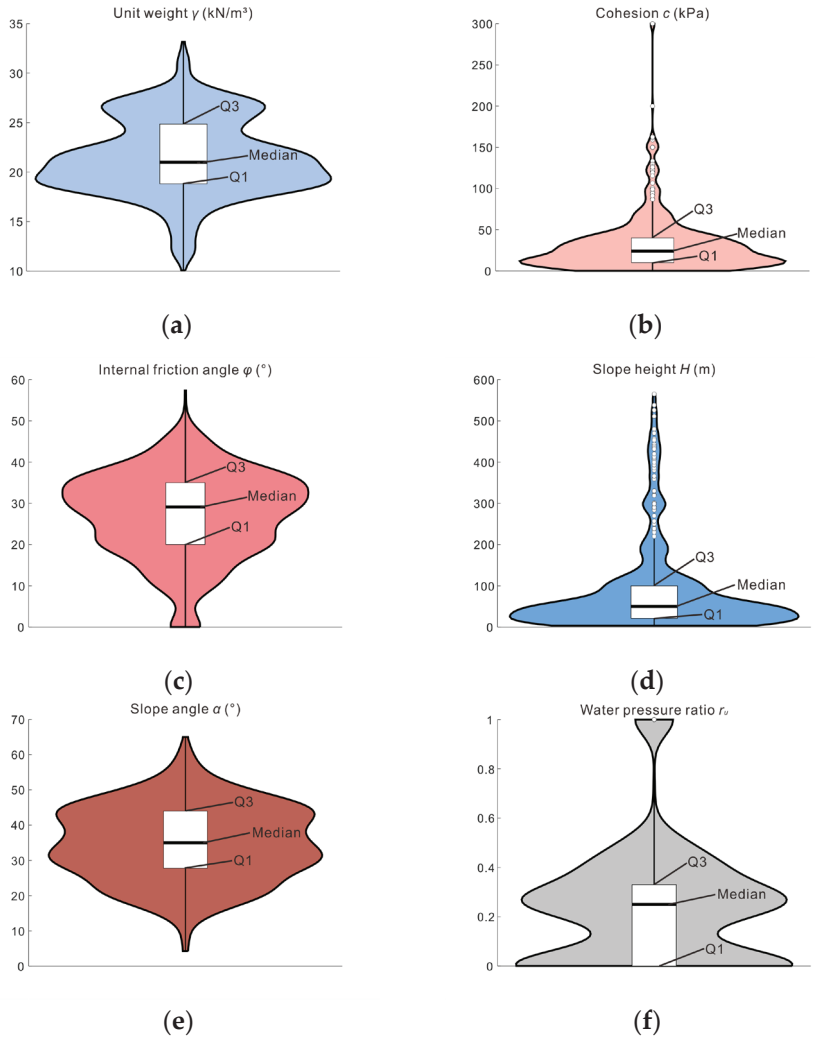


Figure 2. Violin diagram with six input feature parameters: (**a**) Unit weight γ; (**b**) Cohesion c; (**c**) Internal friction angle φ; (**d**) Slope angle α; (**e**) Slope height H and (**f**) Pore pressure ratio r_u.

The violin plot analysis reveals that the pore pressure ratio r_u exhibits a distinctive bimodal distribution with a narrow upper portion and wider lower portion. The median value is 0.3, with the majority of data (interquartile range) concentrated between 0.2 and 0.6. While the distribution spans the full theoretical range (0–1) with values present in nearly all intervals, the data are predominantly clustered around the median. A limited number of outliers were observed without extreme anomalies, indicating a reasonable data distribution pattern.

The violin plots of cohesion and slope height exhibit remarkably similar structures, both displaying a narrow upper part, a wider lower part, and multi-peak patterns in the middle. For slope height, the median value is 50, with the first and third quartiles at 20 and 100, respectively. The data are predominantly distributed between 0 and 200, with numerous outliers present. This indicates that the distribution of slope heights is primarily characterized by low to moderate values, although some exceptionally high slopes exist. Consequently, subsequent numerical modeling designs should focus on low to moderate slopes while accounting for ultra-high slopes.

Regarding cohesion, the median value is 25, with the first and third quartiles at 10 and 45, respectively. The data are mainly concentrated between the quartiles, yet a considerable number of outliers are observed. This suggests that cohesion distribution is dominated by low to moderate values, though a limited number of high cohesion values exist. Thus,

https://doi.org/10.3390/bdcc10030087

numerical simulation schemes should prioritize low to moderate cohesion values while incorporating some cases with exceptionally high cohesion.

The violin plots of slope angle and internal friction angle exhibit a characteristic wide-middle and narrow-ends distribution pattern. Both plots show no outliers, with relatively uniform overall distributions. For the slope angle, the data are predominantly concentrated between 15° and 50°, with neither excessively large nor small values. When the slope angle is too steep, the slope struggles to maintain stability; conversely, excessively gentle slopes exhibit high stability, making landslide occurrences unlikely. Therefore, in subsequent numerical simulation designs, as seen in the data distribution, it is reasonable to assign realistic angles within the 15–50° range to simulations. Regarding the internal friction angle, the data are mainly distributed between 10° and 45°, with no outliers and a relatively uniform spread.

The violin plot of unit weight displays a wide-middle, narrow-ends, and multi-peak distribution. The data are primarily concentrated between 16 and 22, with a secondary cluster between 25 and 28. The overall distribution is centered around the median and quartiles, with no outliers or extreme values observed.

To gain deeper insights into the interrelationships among the parameters, this study constructed both Pearson correlation matrix plots (Figure 3) and paired correlation plots (Figure 4) to further visualize their mutual relationships.

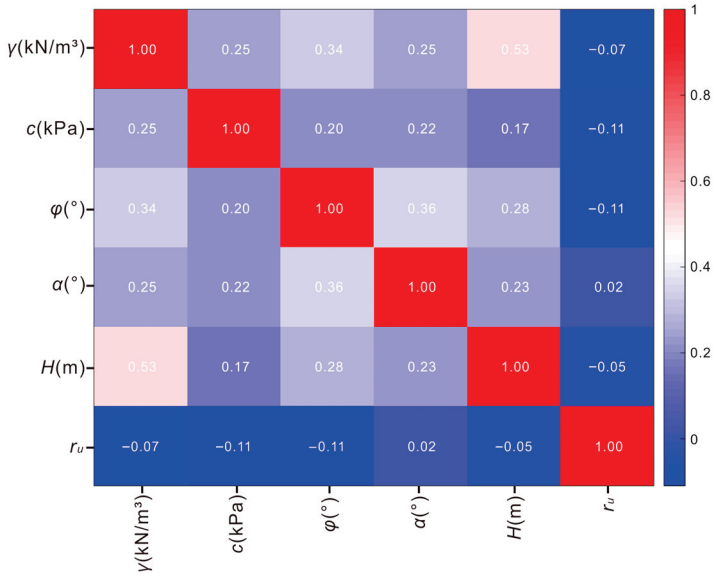


Figure 3. Correlation matrix of six indicators.

As shown in Figure 3, the Pearson correlation matrix clearly demonstrates the correlations between input parameters. The calculation equation for the correlation coefficient is given by Equation (1).

$$r_{XY} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \quad (1)$$

where X and Y represent the parameters, x_i and y_i denote the i observation of the parameters X and Y respectively, $\overline{x}$ and $\overline{y}$ indicate the mean values of parameters Y.

Figure 3 presents the Pearson correlation matrix illustrating inter-parameter correlations. All correlation coefficients remain below 0.55, with the maximum value reaching merely 0.53, indicating weak mutual dependence among the input feature parameters.

https://doi.org/10.3390/bdcc10030087

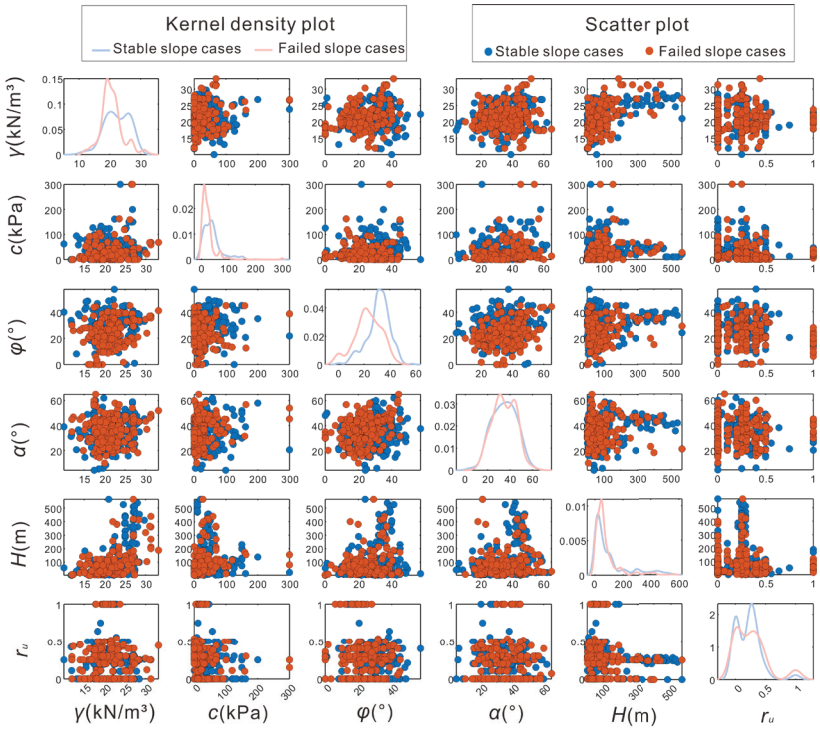


Figure 4. Correlation pair plots of six indicators.

As illustrated in Figure 4, the correlation pair plot feature kernel density plots along the main diagonal and scatter plots in the off-diagonal positions, which visually represent inter-parameter relationships. The kernel density plot along the main diagonal is indicated that, unit weight, internal friction angle, and slope angle approximately follow normal distributions; cohesion and slope height exhibit predominance of low-to-medium values with numerous outliers; while pore pressure ratio displays a multimodal distribution pattern. The off-diagonal scatter plots clearly demonstrate the distribution characteristics of input parameters, showing no significant correlations among parameters, relatively dispersed data distributions, and substantial overlap between stable and unstable cases.

The correlation matrix and paired correlation plots collectively demonstrate weak inter-parameter correlations, indicating that relying on any single feature parameter for prediction would be inadequate. To enhance the accuracy of slope stability assessment, it is essential to incorporate all characteristic parameters as predictive indicators.

3. Methodology

Figure 5 presents a workflow diagram illustrating the proposed methodology for a model establishment method for rapid evaluation of slope stability integrating numerical simulation with engineering cases, as described in this study. Initially, 684 engineering cases were collected through literature mining, with each case comprising seven key parameters: unit weight γ, cohesion c, internal friction angle φ, slope angle α, slope height H, and pore pressure ratio r_u, and slope stability status. These cases collectively constitute the engineering case database. Based on this, statistical analysis was conducted on each parameter in the engineering case database. Building upon these results, numerical simulation schemes were designed according to the maximum values, minimum values, mean values, variances, and distribution densities of the data. Due to the large number of numerical simulations required, MATLAB (version: R2021a) was employed to call Geo-Studio programs for batch computation to achieve rapid calculation. This approach ultimately established the numerical simulation database. The engineering case database

https://doi.org/10.3390/bdcc10030087

was subsequently partitioned, with 80% allocated as the training set and the remaining 20% designated as the test set. The numerical simulation database was then incorporated into the training set to form an enhanced training dataset. Both GRNN and PNN were employed for model training. The trained models were subsequently validated using the test set data, ultimately yielding the slope stability evaluation model. Building upon these findings, we observed that models trained with the augmented dataset incorporating numerical simulations demonstrated superior performance. After systematically evaluating the slope stability assessment models, the optimal model was selected and subsequently validated through precision, recall, and F1-score metrics, confirming its high applicability for practical engineering scenarios.

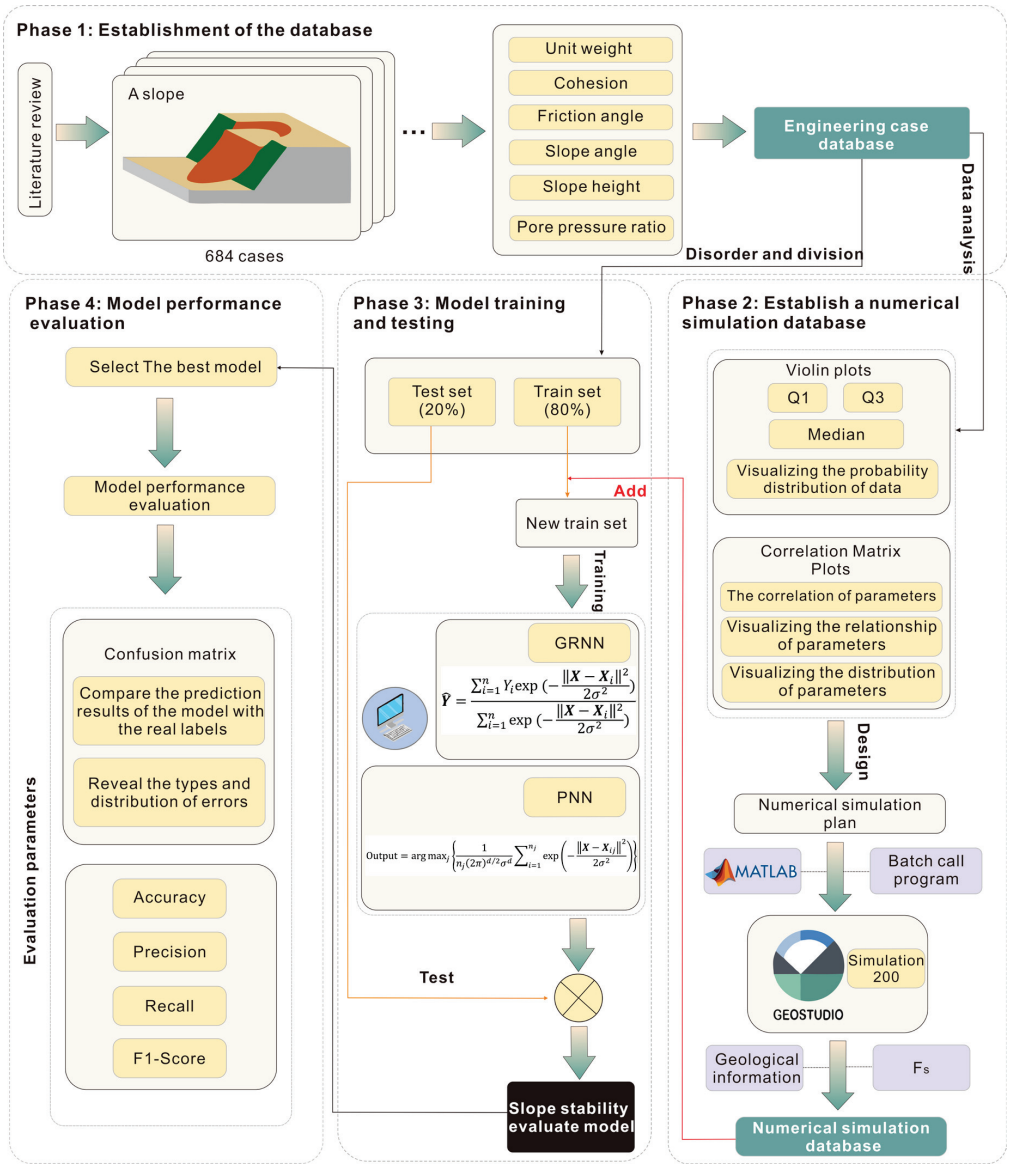


Figure 5. Flowchart of the method for rapidly evaluating slope stability based on engineering case fusion numerical simulation.

3.1. Generalized Regression Neural Network (GRNN)

The GRNN features a four-layer architecture, including an input layer, pattern layer, summation layer, and output layer, with clear division of functions and efficient information transmission.

https://doi.org/10.3390/bdcc10030087

① Input Layer: This layer consists of 7 neurons, corresponding to the six input parameters (unit weight, cohesion, internal friction angle, slope angle, slope height, pore pressure ratio) and the slope stability status, respectively. The input layer directly transmits the received raw data to the pattern layer without any data transformation, ensuring the integrity of the original information.

② Pattern Layer: The number of neurons in this layer is consistent with the number of training samples, totaling 684. Each neuron corresponds to a distinct training sample and uses a radial basis function as its activation function. The specific calculation formula is shown in Equation (2):

$$G_i = \exp\left(-\frac{\|\boldsymbol{X} - \boldsymbol{X}_i\|^2}{2\sigma^2}\right) \tag{2}$$

where $\boldsymbol{X}$ represents the input variables of the network, $\boldsymbol{X}_i$ denotes the learning sample corresponding to the i-th neuron, and σ is the hyperparameter smoothing factor, which plays a key role in regulating the network's fitting effect.

③ Summation Layer: Neurons in this layer perform two types of summation operations to integrate information from the pattern layer.

The first type of neuron conducts arithmetic summation on the outputs of all pattern layer neurons. The connection weight between the pattern layer and this type of neuron is 1, and its transfer function is shown in Equation (4), with the calculation formula as Equation (3):

$$S_1 = \sum_{i=1}^{n} G_i \tag{3}$$

$$f_1(G_i) = G_i \tag{4}$$

The second type of neuron implements weighted summation on the outputs of all pattern layer neurons. The connection weight between the i-th neuron in the pattern layer and this type of neuron is the first element of the i-th output sample $\boldsymbol{Y}_i$, and the transfer function is given by Equation (6), with the calculation formula as Equation (5):

$$S_2 = \sum_{i=1}^{n} Y_i G_i \tag{5}$$

$$f_2(G_i) = Y_i G_i \tag{6}$$

④ Output Layer: This layer contains 1 neuron, which normalizes the two types of summation results from the summation layer (i.e., divides the weighted summation result by the arithmetic summation result) to obtain the final estimation of slope stability. The calculation formula is shown in Equation (7):

$$\hat{\boldsymbol{Y}} = \frac{S_2}{S_1} = \frac{\sum_{i=1}^{n} Y_i \exp\left(-\frac{\|\boldsymbol{X} - \boldsymbol{X}_i\|^2}{2\sigma^2}\right)}{\sum_{i=1}^{n} \exp\left(-\frac{\|\boldsymbol{X} - \boldsymbol{X}_i\|^2}{2\sigma^2}\right)} \tag{7}$$

3.2. Probabilistic Neural Network (PNN)

The PNN also adopts a four-layer structure, namely the input layer, pattern layer, summation layer, and output layer, which is similar to GRNN but differs in core calculation logic and functional positioning.

① Input Layer: Composed of 7 neurons, it corresponds to the six characteristic parameters and the slope stability status, consistent with the input layer of GRNN. The

https://doi.org/10.3390/bdcc10030087

input layer directly forwards the received data to the pattern layer without additional processing, maintaining the consistency of data input format.

② Pattern Layer: There are 684 neurons in this layer, each corresponding to a training sample. The activation function of the neurons is a radial basis function, and the calculation formula is shown in Equation (8):

$$G_i = \exp\left(-\frac{\|\boldsymbol{X} - \boldsymbol{X}_i\|^2}{2\sigma^2}\right) \tag{8}$$

where $\boldsymbol{X}$ represents the input variables of the network, and $\boldsymbol{X}_i$ denotes the learning sample corresponding to the i-th neuron, which is used to calculate the similarity between the input data and the training samples.

③ Summation Layer: Each neuron in this layer is associated only with pattern layer neurons of the same category, and based on the Parzen window function method, it estimates the conditional probability density of different categories. The specific calculation formula of the Parzen window probability density estimation is shown in Equation (9):

$$p_j(\boldsymbol{X}) = \frac{1}{n_j(2\pi)^{d/2}\sigma^d}\sum_{i=1}^{n_j}\exp\left(-\frac{\|\boldsymbol{X} - \boldsymbol{X}_{ij}\|^2}{2\sigma^2}\right) \tag{9}$$

where d denotes the number of input nodes, n_j is the number of training samples in the j-th category of the input layer, σ is the hyperparameter of PNN (which affects the smoothness of the probability density curve), $\boldsymbol{X}$ represents the input vector, and $\boldsymbol{X}_{ij}$ denotes the input vector of the i-th training sample in the j-th category.

④ Output Layer: This layer includes 1 neuron, which determines the final classification result through a competitive mechanism. It compares the conditional probability density values $p_j(\boldsymbol{X})$ of different categories obtained by the summation layer and selects the category with the highest probability as the output of slope stability status (stable or unstable), i.e.,

$$\text{Output} = \text{argmax}_j\{p_j(\boldsymbol{X})\} \tag{10}$$

3.3. Verification and Validation

This study developed MATLAB programs for the GRNN and PNN models to implement the slope stability evaluation method integrating engineering case and numerical simulation databases. The core program framework and basic algorithm logic of the GRNN/PNN models are derived directly from the mature, verified code provided by the author of *30 Case Studies of Intelligent Algorithms in MATLAB* [31], a widely recognized reference work in the field of MATLAB-based intelligent algorithm development. Given that the foundational code has been rigorously validated and applied in numerous academic and engineering scenarios as documented in the reference book, the core algorithm logic of GRNN/PNN does not require re-verification for correctness. Only scenario-specific customizations were made to adapt the code to slope stability prediction scenarios by the MATLAB R2024a, with no modifications to the core algorithm logic.

This study proposes a rapid modeling method for slope stability evaluation by fusing engineering case databases with numerical simulation databases, and a dedicated MATLAB program was developed to implement this method. The program development process began with data preprocessing, where raw data were converted into machine-readable formats to ensure compatibility with neural network modeling. In this research, the parameterization of input and output variables for the models was defined as follows:

https://doi.org/10.3390/bdcc10030087

(1) Input variables: Slope angle, unit weight, slope height, internal friction angle, cohesion, and pore pressure ratio, all directly adopted from the collected engineering and numerical simulation data;
(2) Output variable (slope stability status): Categorically defined as 1 for stable slopes and 2 for unstable (failed) slopes.

Subsequently, both GRNN and PNN models were constructed in MATLAB using the preprocessed dataset, which was randomly partitioned into a training subset (80%) and a testing subset (20%). The six slope characteristic parameters were designated as the input variables for the neural networks, while slope stability status served as the output variable. Based on different combinations of these input feature parameters, the neural networks generated a total of 63 slope stability evaluation models, with the specific input parameter configuration for each model detailed in Figure 6.

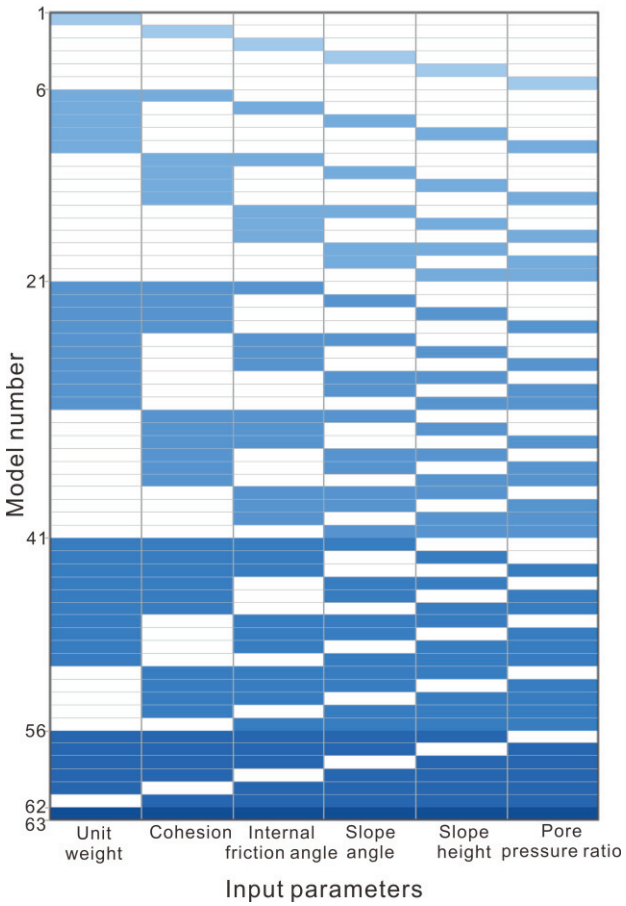


Figure 6. The input parameters corresponding to each model. Keys: No. 1–6 are single-input models, No. 7–21 are two-input models, No. 22–41 are three-input models, No. 42–56 are four-input models, No. 57–62 are five-input models, and No. 63 is a six-input model.

3.4. Model Assessment

Comprehensive performance evaluation was conducted for all successfully trained models using classification metrics including accuracy, precision, recall, and F1-Score, through which the optimal intelligent slope stability identification model was determined. The evaluation of predictive performance using the aforementioned metrics requires pre-

https://doi.org/10.3390/bdcc10030087

generation of the model's confusion matrix. As this study deals with a binary classification problem, the resulting confusion matrix is a 2 × 2 matrix that can be represented as:

$$M = \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{pmatrix} \tag{11}$$

In the confusion matrix, the main diagonal elements x_{11}, x_{22} represent the number of correctly predicted samples by the model. The first row of the matrix indicates samples classified as Category 1, while the second row represents samples classified as Category 2. The first column of the matrix represents all samples predicted as Category 1, while the second column denotes those predicted as Category 2. For example, [element] represents the number of samples that actually belong to category 1 (slope stability) but were incorrectly predicted by the model as category 2 (slope instability).

The discriminative capability of the model can be evaluated based on performance metrics.

Accuracy: The proportion of correctly classified samples to the total number of samples, which refers to the ratio of accurately identified slope stability cases to the total identified samples, is calculated by the following Equation (12):

$$Accuracy = \left(\frac{1}{n} \sum_{i=1}^{m} x_{ii} \right) \times 100\% \tag{12}$$

Precision: defined as the proportion of correctly predicted slope stability cases among all samples predicted as that stability category, is calculated by the following Equation (13):

$$Precision = \left(\frac{x_{ii}}{x_{+i}} \right) \times 100\%, \;\; i = 1, 2 \tag{13}$$

Recall: defined as the proportion of correctly predicted slope stability cases among all actual cases of that stability category, is computed using the following Equation (14):

$$Recall = \left(\frac{x_{ii}}{x_{i+}} \right) \times 100\%, \;\; i = 1, 2 \tag{14}$$

F1-Score: defined as the harmonic mean of precision and recall that balances these two metrics, is calculated by the following Equation (15):

$$F1 - Score = \frac{2 \times (Precision \times Recall)}{(Precision + Recall)} \tag{15}$$

where n represents the total number of samples in the dataset and m denotes the total number of categories, which is set to 2 in this study; x_{+i} represents the actual number of samples belonging to category i and x_{i+} denotes the predicted number of samples classified as category i.

3.5. Adjustment of Neural Network Parameters

In the parameter thresholds of hidden layer neurons in neural networks, the spread value governs the expansion rate of the radial basis function. It is a critical hyperparameter affecting the modeling performance of both GRNN and PNN. The default value of this parameter is set to 0.1 for both network architectures. Notably, the spread value exhibits a dual influence on model behavior. A larger spread value yields smoother function fitting but concurrently increases approximation error and elevates computational complexity. Conversely, a smaller spread value enhances the precision of function approximation yet compromises fitting smoothness, thereby introducing a heightened risk of model overfitting. Consequently, the rational selection of the spread value is pivotal to ensuring the reliability

https://doi.org/10.3390/bdcc10030087

of neural network simulation results. The core principle for parameter selection involves balancing two objectives. The spread value should be sufficiently large to cover a broad input range, ensuring effective activation of all neurons. Meanwhile, it must not be excessively large to prevent overfitting, ultimately achieving a relatively optimal trade-off.

To strictly comply with the fundamental principle of hyperparameter tuning independent of the test set in machine learning, this study adopted a repeated random partitioning validation strategy based on the entire engineering case database. The detailed implementation process is outlined as follows:

1. The complete engineering case database was designated as the base dataset for hyperparameter optimization. For each candidate spread value, the base dataset was randomly split into a training subset (80%) and a validation subset (20%). Independent partitioning was performed for each individual trial to avoid systematic bias.
2. The search space for the spread value was defined as 0.1 to 4.0 with a step size of 0.01. For each spread value within this range, 10 independent trials were conducted. In each trial, the model was trained on the newly partitioned training subset and its performance was evaluated on the corresponding validation subset.
3. To mitigate the impact of randomness inherent in a single data partitioning, the average validation accuracy of each spread value was computed across the 10 repeated trials, providing a robust measure of parameter performance.
4. The spread value corresponding to the highest overall average validation accuracy was selected as the optimal hyperparameter. Importantly, this selection process was conducted without any reference to the test set data, ensuring the strict independence of hyperparameter tuning and the subsequent final model evaluation.
5. Subsequent to determining the optimal spread value, the entire engineering case database was re-partitioned into a formal training set (80%) and a test set (20%). This unique partitioning scheme was exclusively used for model training and performance assessment in the subsequent sections of this study.

As illustrated in Figure 7, GRNN consistently outperformed PNN in terms of average validation accuracy across the entire range of spread values, while both networks exhibited analogous variation trends. Within the spread value range of 0.1 to 0.6, the average validation accuracy of GRNN and PNN attained their respective peaks, reaching 85% and 83%. When the spread value exceeded 0.6, the validation accuracy of both models initially remained stable and subsequently exhibited a gradual declining trend. Considering the comprehensive performance metrics of validation accuracy and model generalization ability, the optimal spread value was ultimately determined to be 0.5 in this study.

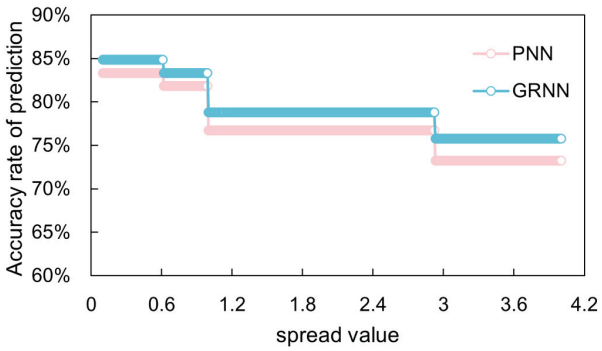


Figure 7. Analysis of the Influence of spread Value on Prediction Accuracy.

https://doi.org/10.3390/bdcc10030087

3.6. Numerical Simulation Database

This study proposes a model establishment method for the rapid evaluation of slope stability, integrating numerical simulation with engineering cases. Given the limited scale of the engineering case database, we expanded the database by designing numerical simulation schemes based on the distribution characteristics of six key parameters derived from existing cases. The specific procedure is as follows: based on the original database compiled through literature mining, we first analyzed the statistical characteristics of each parameter, including maximum value, minimum value, median, distribution density, mean, and variance. On this basis, we assumed that the parameters follow a normal distribution and conducted random sampling. During sampling, boundary constraints were imposed on the parameter values to ensure that the sampled maximum and minimum values did not exceed the corresponding ranges of the original database. In addition, the sampling results were iteratively calibrated to make the sampling database match the original database as closely as possible in terms of median, mean, variance, and distribution density. Accordingly, a candidate database for numerical simulation was established. Subsequently, numerical simulations were performed for each candidate sample, and the corresponding factor of safety F_s was calculated. Samples for which the factor of safety could be obtained were regarded as reasonable and valid and thus retained, while those that failed to yield the factor of safety were considered unreasonable and excluded [20]. Ultimately, a simulation database consisting of 200 valid numerical simulation cases was constructed.

The selection of 200 numerical simulation cases is based on the core purpose of this study, which is using the simulation database to supplement the real-case database. If an excessive number of simulated samples were included, the model might be more prone to learning the structural assumptions or generation biases implicit in the simulation mechanism, thereby undermining its generalization ability to real-world data. Therefore, 200 simulated samples were adopted as the initial setting for validation-oriented analysis.

In this study, slope stability analysis was conducted using GeoStudio 2012 (developed by GEO-SLOPE International Ltd., Calgary, AB, Canada), specifically utilizing its SLOPE/W module for limit equilibrium analysis [32]. The input parameters for numerical simulations include unit weight γ, cohesion c, internal friction angle φ, slope angle α, slope height H, and pore pressure ratio r_u. SLOPE/W incorporates the Mohr–Coulomb constitutive model expressed in Equation (16). Thus, only cohesion and internal friction angle are required as strength parameters, while unit weight is directly specified as a material property. Slope height and slope angle are used to define the slope geometry, and the slope profile is constructed accordingly during model setup. The pore pressure ratio is defined as the ratio of pore-water pressure at a specific point in the slope to the corresponding vertical stress. In this study, the initial pore-water pressure field was obtained by converting the preset pore pressure ratio, and the factor of safety for each case was then calculated using the simplified Bishop method expressed in Equation (17). In accordance with the relevant specification (GB 50330-2013), cases with a factor of safety greater than 1.15 are classified as stable, while those with a factor of safety less than 1.15 are classified as unstable [33]. This process ultimately generated the numerical simulation database.

$$\tau = c + \sigma tan\phi \tag{16}$$

where τ denotes the shear strength at failure, c denotes cohesion, σ denotes the effective normal stress, and ϕ denotes the internal friction angle.

https://doi.org/10.3390/bdcc10030087

$$F_s = \frac{\sum_{i=1}^{n}\left[\frac{cb_i+(W_i-u_ib_i)\tan\phi}{1+\frac{\tan\phi\tan\alpha_i}{F_s}}\right]}{\sum_{i=1}^{n} W_i\sin\alpha_i} \tag{17}$$

where F_s denotes the factor of safety, b_i denotes the width of slice i, W_i denotes the weight of slice i, u_i denotes the pore-water pressure of slice i, α_i denotes the inclination angle of slice i, and n denotes the number of slices.

As illustrated in Figure 8, parallel coordinate plots are used to visualize the obtained numerical simulation database containing 200 datasets, where each polyline represents an individual dataset. The left axis corresponds to stable slopes, accounting for 79 cases, while the right axis corresponds to unstable slopes, totaling 121 cases.

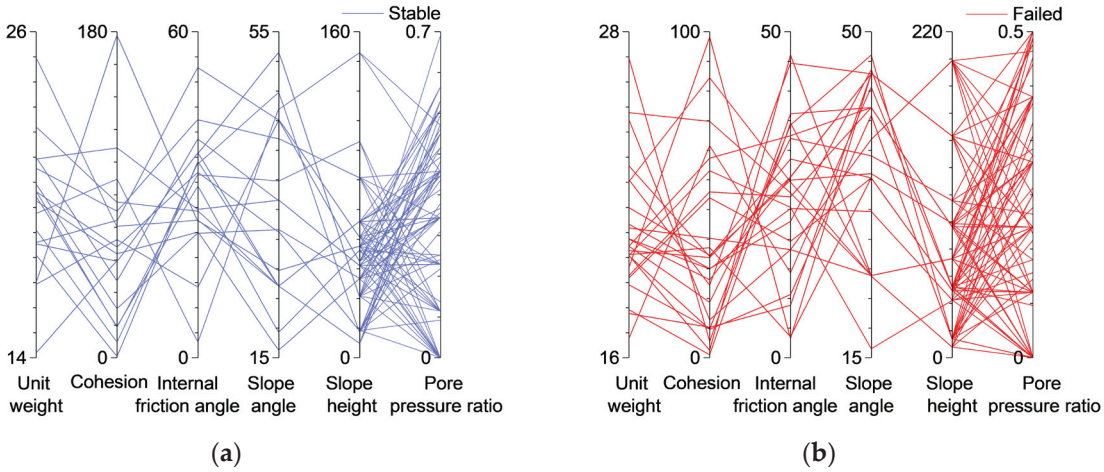


Figure 8. Numerical simulation database: (**a**) Cases of stable slope; (**b**) Cases of failed slope.

4. Results

4.1. Analysis of Experimental Results

The process begins by training and testing the neural network using data from the engineering case database, with 80% of the data allocated to the training set and the remaining 20% reserved for the testing set. The test results of GRNN are presented in Figure 9. The average prediction accuracy shows an increasing trend with the addition of input parameters. Initial observations indicate that the model exhibits relatively poor predictive performance when fewer parameters are incorporated. However, a significant improvement in predictive capability is observed when the number of input parameters exceeds three. Notably, Model 57 and Model 63 demonstrate identical prediction accuracy (85%), suggesting that including the pore pressure ratio as an additional input parameter does not lead to substantial performance enhancement.

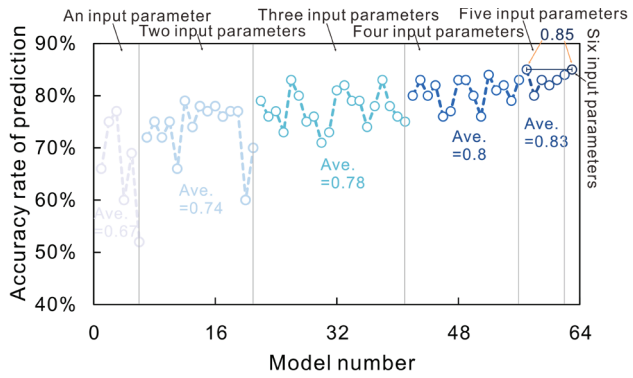


Figure 9. The test result graph of GRNN.

https://doi.org/10.3390/bdcc10030087

The PNN results are shown in Figure 10. In terms of overall prediction performance, the model with the highest accuracy utilizes only two input parameters. The models with fewer input parameters demonstrated higher prediction accuracy compared to those requiring more input parameters. Moreover, the PNN exhibited a distinct trend where prediction accuracy initially increased but subsequently decreased as the number of input parameters grew. Furthermore, comparative analysis reveals that Model 57 outperforms Model 63, unexpectedly indicating that incorporating the pore pressure ratio as an input parameter actually reduces prediction accuracy.

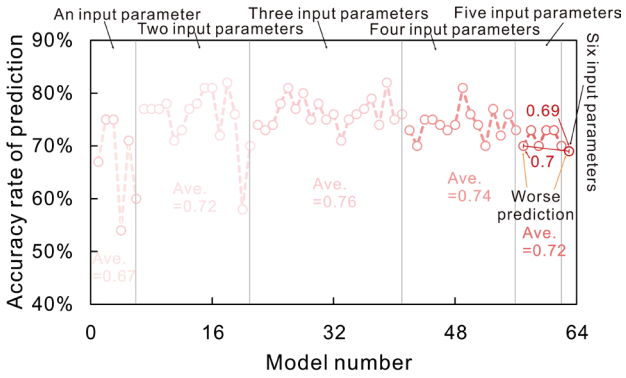


Figure 10. The test result graph of PNN.

The experimental results demonstrate that both GRNN and PNN exhibit low sensitivity to the pore pressure ratio parameter. Notably, when the pore pressure ratio is included as an input feature in PNN, the prediction accuracy shows a decreasing trend instead. As the only external triggering factor among the input parameters, the pore pressure ratio quantifies the proportion of pore water pressure relative to total stress in soil, directly reflecting the weakening effect of water on soil strength. Moreover, slope instability typically exhibits dynamic evolutionary characteristics. The pore pressure ratio can effectively capture critical real-time variations during this process, enabling the model to accurately identify the temporal evolution pattern where rising pressure leads to strength degradation and ultimately triggers instability. Without this parameter, the model may rely solely on static features (e.g., slope angle, unit weight) and fail to accurately predict dynamically induced landslides. Therefore, the pore pressure ratio is crucial for predictive models. To address its limited contribution to prediction accuracy, we integrated the numerical simulation database into the neural network training set (Figure 5). The GRNN was subsequently re-evaluated on the test set, with the updated results presented in Figure 11.

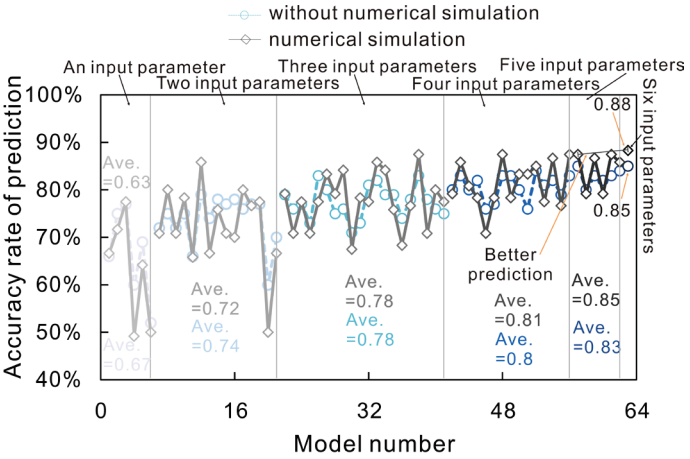


Figure 11. Comparative test results of GRNN with/without numerical simulation database incorporation.

https://doi.org/10.3390/bdcc10030087

As shown in Figure 11, under identical test set conditions, we trained the GRNN using two distinct training datasets: the original training set and an augmented training set integrated with the numerical simulation database. The experimental results demonstrate that incorporating numerical simulation data into the training set reduces model prediction performance when fewer parameters are used. However, when the number of parameters exceeds three, the model's predictive capability improves significantly, with a notable increase in average accuracy. Remarkably, the enhanced training set achieves peak prediction accuracy (88.3%) when all input parameters are included. Furthermore, the average prediction accuracy exhibits a consistent upward trend as the number of parameters increases.

The experimental results in Figure 12 further validate the performance enhancement effect of the numerical simulation database on the model. It can be observed that when fewer than 4 parameters are used, there is no significant difference in prediction performance before and after incorporating the numerical simulation database into the training set. However, for cases involving more than 4 parameters, the inclusion of the numerical simulation database in the neural network's training set demonstrates significant performance improvement. Similar to GRNN, the model achieves peak prediction accuracy (88.3%) when all parameters are incorporated. Regarding PNN's performance trend, the enhanced training set exhibits a rising prediction accuracy trend with increasing parameters, showing strong consistency with GRNN's prediction results.

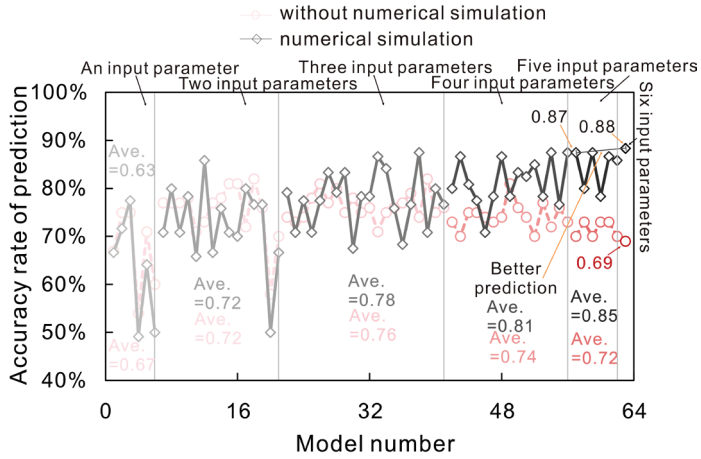


Figure 12. Comparative test results of PNN with/without numerical simulation database incorporation.

The experimental results demonstrate that incorporating the numerical simulation database into the neural network training set, followed by model retesting, can significantly enhance prediction accuracy.

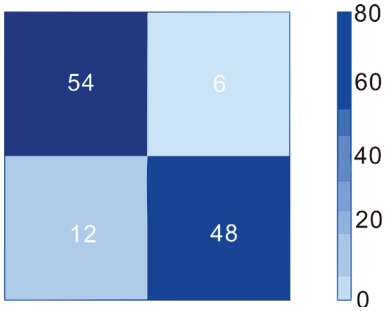
4.2. Performance Evaluation of the Optimal Model

Based on the comprehensive model evaluation results, the GRNN neural network achieves optimal prediction accuracy when incorporating the following six input parameters: unit weight, cohesion, internal friction angle, slope angle, slope height, and pore pressure ratio. To evaluate the feasibility of the optimal neural network with selected parameter combinations, the performance metrics of GRNN were computed before and after incorporating the numerical simulation database into the training set, as illustrated in the following.

As can be seen from Figures 13 and 14, after introducing the numerical simulation database into the training set, both the recall rate and precision of the model have significantly increased. When the numerical simulation database was not introduced, the F1-Score

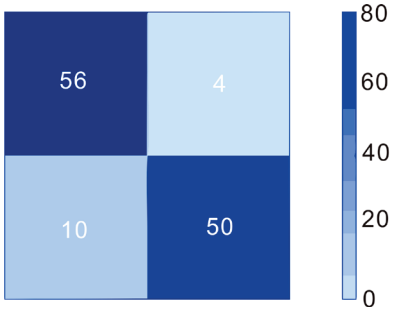
https://doi.org/10.3390/bdcc10030087

of the model was 0.85, while after introducing the numerical simulation database, the F1-Score of the model was 0.88. In summary, the incorporation of the numerical simulation database has led to a marked improvement in model performance.



	Precision	Recall
Failure	81.8%	90%
Stable	88.9%	80%

Figure 13. The confusion matrix of the numerical simulation database has not been added.



	Precision	Recall
Failure	84.8%	93.3%
Stable	92.6%	83.3%

Figure 14. Adding the confusion matrix of the numerical simulation database.

5. Discussion

This study proposes a model establishment method for the rapid evaluation of slope stability, integrating numerical simulation with engineering cases. The methodology commenced with the collection of 684 engineering case datasets via literature mining. In machine learning, fundamental data validity checks typically encompass the detection of missing values, outliers, duplicate entries, and verification of data consistency. For this study, since the data were derived from published literature, only reproducibility verification was deemed necessary. Subsequent data analysis involved examining data distributions using violin plots, as well as constructing correlation matrices and pairwise correlation plots to assess inter-variable relationships.

To address the limited significance of the pore pressure ratio as an input feature parameter and further improve the model's predictive performance, this study incorporates a numerical simulation database into the training set. The enhanced predictive performance achieved through the numerical simulation database primarily stems from its superior data quality compared to engineering case data, which is characterized by higher validity, fewer outliers, and reduced extreme values. These attributes enable neural networks to more effectively learn and recognize underlying patterns [34]. Additionally, the numerical simulation database was deliberately incorporated into the neural network's training set rather than the test set, as the simulated data were specifically designed based on engineering case data and adhere to established physical principles, rather than representing unknown samples from real-world scenarios [30]. Thus, the integration of numerical simulation data into the training set serves as a complementary enhancement to the original engineering case database.

This study focuses on the application of GRNN and PNN to slope stability classification. In recent years, the rapid advancement of machine learning has fueled the adoption

https://doi.org/10.3390/bdcc10030087

of a growing array of algorithms for slope stability prediction. However, the scope of the present research is confined to the utilization of neural network-based models for this classification task. Future investigations should prioritize two primary directions. The first direction involves exploring a broader range of machine learning approaches, including classical algorithms such as random forests and decision trees. The second direction centers on developing ensemble learning models integrated with multi-model collaboration schemes. While standalone models often demonstrate limited adaptability to the complex nonlinear data characteristics and abrupt condition changes inherent in slope stability analysis, ensemble methods can effectively integrate the complementary strengths of diverse algorithms, thereby substantially enhancing the predictive performance and robustness of classification models.

The proposed model establishment method for rapid slope stability evaluation, which integrates numerical simulation with engineering cases, involves the construction and quantitative determination of numerical simulation databases. In this study, a preliminary dataset consisting of 200 sets of numerical simulation data was used to improve the model's predictive accuracy. Since the test set employed to evaluate model performance was a subset of the real-world database, the model's performance on entirely new and mutually independent datasets remains to be verified. Future work should incorporate external datasets from independent sources that may have different distributions for testing and comparative analyses, to more comprehensively assess and demonstrate the model's generalization capability and practical applicability.

Furthermore, future research will focus on determining the optimal size of the numerical simulation database and developing more rational numerical simulation schemes. First, the optimal size of the numerical simulation database can be determined by investigating the ratio between the numerical simulation database and the engineering case database. By examining how different volumes of engineering case data correspond to the optimal amount of numerical simulation data, the most suitable size of the numerical simulation database can be identified. Finally, regarding the design of numerical simulation schemes, future studies should carefully consider both data distribution density and the number of extreme values/outliers when formulating numerical simulation strategies. This approach ensures that the developed simulation schemes retain the patterns observed in the original engineering case data.

6. Conclusions

Through comprehensive literature mining, this study collected a total of 684 engineering cases to establish an engineering case database. However, due to limitations such as insufficient data coverage and the inability to represent all possible slope states, we innovatively proposed a statistically based design method for numerical simulation. Based on this approach, a numerical simulation database containing 200 cases was constructed. By incorporating these 200 simulated datasets into the neural network's training set while keeping the test set unchanged, the accuracy of the slope stability evaluation model was significantly improved. Ultimately, this study developed a novel modeling approach for rapid slope stability assessment by effectively integrating numerical simulations with real-world engineering cases. This study innovatively integrates numerical simulation data with engineering case data, providing a novel approach to enhance the prediction accuracy of slope stability. Future research will focus on optimizing the design methodology of numerical simulation databases, expanding the scale of simulated datasets, and employing more machine learning algorithms for comprehensive analysis.

https://doi.org/10.3390/bdcc10030087

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/bdcc10030087/s1.

Author Contributions: Conceptualization, J.J. and M.C.; Methodology, J.J. and D.L.; Software, Q.Y.; Validation, D.L.; Formal analysis, D.L., L.W. and W.Z.; Investigation, D.L., Q.Y., Z.Z., L.W. and W.Z.; Resources, Z.Z.; Data curation, J.J. and Q.Y.; Writing—original draft preparation, J.J. and L.W.; Writing—review and editing, M.C.; Visualization, L.W.; Supervision, J.J., Z.Z., W.Z. and M.C.; Project administration, J.J.; Funding acquisition, M.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (52509143), the Anhui Provincial Natural Science Foundation (No. 2408085QE154), and Open Fund Research at the State Key Laboratory of Hydraulics and Mountain River Engineering (SKHL2421). These supports are gratefully acknowledged.

Data Availability Statement: Data will be made available on request.

Conflicts of Interest: Author J.J., D.L. and Q.Y. were employed by the company [Shandong Electric Power Engineering Consulting Institute Corp., Ltd.]. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Tao, T.; Shi, W.; Liang, F. Failure mechanism and evolution of the Jinhaihu landslide in Bijie City, China, on January 3, 2022. *Landslides* **2022**, *19*, 2727–2736. [CrossRef]
2. Ma, S.; Xu, C.; Xu, X. Characteristics and causes of the landslide on July 23, 2019 in Shuicheng, Guizhou Province, China. *Landslides* **2020**, *17*, 1441–1452. [CrossRef]
3. Zhang, X.; Zhang, S. Research on the Path of Constructive News of Disaster Report in the Era of Convergence-Media—Taking the Coal Mine Collapse Accident in Alxa Left Banner, Inner Mongolia as an Example. *Acad. J. Humanit. Soc. Sci.* **2023**, *6*, 116–122.
4. Redshaw, P.; Boon, D.P.; Campbell, G.; Willis, M.G.; Mattai, J.; Free, M.; Jordan, C.J.; Kemp, S.J.; Morley, A.; Thomas, M.F. The 2017 Regent Landslide, Freetown Peninsula, Sierra Leone. *Q. J. Eng. Geol. Hydrogeol.* **2019**, *52*, 435–444. [CrossRef]
5. Samui, P.; Kothari, D.P. Utilization of a least square support vector machine (LSSVM) for slope stability analysis. *Sci. Iran.* **2011**, *18*, 53–58. [CrossRef]
6. Gordan, B.; Jahed Armaghani, D.; Hajihassani, M. Prediction of seismic slope stability through combination of particle swarm optimization and neural network. *Eng. Comput.* **2016**, *32*, 85–97.
7. Gao, W.; Raftari, M.; Rashid, A.S.A. A predictive model based on an optimized ANN combined with ICA for predicting the stability of slopes. *Eng. Comput.* **2020**, *36*, 325–344. [CrossRef]
8. Zhou, J.; Li, E.; Yang, S.; Wang, M.; Shi, X.; Yao, S.; Mitri, H.S. Slope stability prediction for circular mode failure using gradient boosting machine approach based on an updated database of case histories. *Saf. Sci.* **2019**, *118*, 505–518. [CrossRef]
9. Bui, X.N.; Nguyen, H.; Choi, Y. Prediction of slope failure in open-pit mines using a novel hybrid artificial intelligence model based on decision tree and evolution algorithm. *Sci. Rep.* **2020**, *10*, 9939. [CrossRef]
10. Deng, C.; Hu, H.; Zhang, T. Rock slope stability analysis and charts based on hybrid online sequential extreme learning machine model. *Earth Sci. Inform.* **2020**, *13*, 729–746.
11. Moayedi, H.; Osouli, A.; Nguyen, H. A novel Harris hawks' optimization and k-fold cross-validation predicting slope stability. *Eng. Comput.* **2021**, *37*, 369–379.
12. Sun, J.; Wu, S.; Zhang, H.; Zhang, X.; Wang, T. Based on multi-algorithm hybrid method to predict the slope safety factor-stacking ensemble learning with bayesian optimization. *J. Comput. Sci.* **2022**, *59*, 101587. [CrossRef]
13. Zhang, W.; Li, H.; Han, L.; Chen, L.; Wang, L. Slope stability prediction using ensemble learning techniques: A case study in Yunyang County, Chongqing, China. *J. Rock Mech. Geotech. Eng.* **2022**, *14*, 1089–1099. [CrossRef]
14. Yang, H.Q.; Shi, C.; Zhang, L. Ensemble Learning of Soil–Water Characteristic Curve for Unsaturated Seepage Using Physics-Informed Neural Networks. *Soils Found.* **2025**, *65*, 101556. [CrossRef]
15. Gao, W. The Application of Machine Learning in Geotechnical Engineering. *Appl. Sci.* **2024**, *14*, 4712. [CrossRef]
16. Ma, N.; Zhang, Y.; Yao, Z. Slope Stability Prediction Under Seismic Loading Based on the EO-LightGBM Algorithm. *Front. Earth Sci.* **2025**, *13*, 1591219.
17. Li, X.; Nishio, M.; Sugawara, K.; Iwanaga, S.; Chun, P.-j. Surrogate Model Development for Slope Stability Analysis Using Machine Learning. *Sustainability* **2023**, *15*, 10793. [CrossRef]

https://doi.org/10.3390/bdcc10030087

18. Khan, M.I.; Fei, J.; Chen, X.; Hamza, M. Development of Machine Learning Surrogate Models for Slope Stability Prediction Using AI Techniques: A Case Study of the Meizhou Landslide. *Sci. Rep.* **2026**, *16*, 3362. [CrossRef]
19. Hsiao, C.-H.; Chen, A.Y.; Ge, L.; Yeh, F.-H. Performance of Artificial Neural Network and Convolutional Neural Network on Slope Failure Prediction Using Data from the Random Finite Element Method. *Acta Geotech.* **2022**, *17*, 5801–5811. [CrossRef]
20. Liu, L.; Zhao, G.; Liang, W. Slope Stability Prediction Using k-NN-Based Optimum-Path Forest Approach. *Mathematics* **2023**, *11*, 3071.
21. Ma, J.; Jiang, S.; Liu, Z.; Ren, Z.; Lei, D.; Tan, C.; Guo, H. Machine Learning Models for Slope Stability Classification of Circular Mode Failure: An Updated Database and Automated Machine Learning (AutoML) Approach. *Sensors* **2022**, *22*, 9166. [CrossRef]
22. Huang, F.; Xiong, H.; Chen, S.; Lv, Z.; Huang, J.; Chang, Z.; Catani, F. Slope Stability Prediction Based on a Long Short-Term Memory Neural Network: Comparisons with Convolutional Neural Networks, Support Vector Machines and Random Forest Models. *Int. J. Coal Sci. Technol.* **2023**, *10*, 18. [CrossRef]
23. Jing, Y.; Li, Y.; Chang, J.; Liu, Z.; Ni, Z.; Wang, Q.; Gao, D. Factor of Safety Prediction for Slope Stability Using PCA and BPNN in Guangdong's H Mining Area. *Sci. Rep.* **2025**, *15*, 12804. [CrossRef]
24. Shiferaw, H.M. Study on the influence of slope height and angle on the factor of safety and shape of failure of slopes based on strength reduction method of analysis. *Beni-Suef Univ. J. Basic Appl. Sci.* **2021**, *10*, 31. [CrossRef]
25. Selmi, M.; Baroth, J.; Ltifi, M.; Hassis, H. Reliability analyses of slope stability. *Eur. J. Environ. Civ. Eng.* **2010**, *14*, 1227–1257.
26. Bishop, A.W.; Morgenstern, N.R. Stability coefficients for earth slopes. *Geotechnique* **1960**, *10*, 129–150. [CrossRef]
27. Michalowski, R.L. Slope stability analysis: A kinematical approach. *Géotechnique* **1995**, *45*, 283. [CrossRef]
28. Gelisli, K.; Kaya, T.; Babacan, A.E. Assessing the factor of safety using an artificial neural network: Case studies on landslides in Giresun, Turkey. *Environ. Earth Sci.* **2015**, *73*, 8639–8646. [CrossRef]
29. Kumar, S.; Choudhary, S.S.; Burman, A. The effect of slope height and angle on the safety factor and modes of failure of 3D slopes analysis using limit equilibrium method. *Beni-Suef Univ. J. Basic Appl. Sci.* **2023**, *12*, 84. [CrossRef]
30. Karir, D.; Ray, A.; Bharati, A.K.; Chaturvedi, U.; Rai, R.; Khandelwal, M. Stability prediction of a natural and man-made slope using various machine learning algorithms. *Transp. Geotech.* **2022**, *34*, 100745. [CrossRef]
31. Yu, L.; Shi, F.; Wang, H.; Hu, F. *30 Case Studies of Intelligent Algorithms in MATLAB*, 2nd ed.; Beihang University Press: Beijing, China, 2011. (In Chinese)
32. GEO-SLOPE International Ltd. *Stability Modeling with SLOPE/W*; GEO-SLOPE International Ltd.: Calgary, AB, Canada, 2012.
33. *GB 50330-2013*; Technical Code for Building Slope Engineering. China Architecture & Building Press: Beijing, China, 2014.
34. Yadav, D.K.; Chattopadhyay, S.; Tripathy, D.P. Enhanced slope stability prediction using ensemble machine learning techniques. *Sci. Rep.* **2025**, *15*, 7302. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

https://doi.org/10.3390/bdcc10030087

fire

MDPI

Article

Inferring Wildfire Ignition Causes in Spain Using Machine Learning and Explainable AI

Clara Ochoa [1,*], Magí Franquesa [2,3], Marcos Rodrigues [4,5] and Emilio Chuvieco [1]

[1] Environmental Remote Sensing Research Group, Department of Geology, Geography and the Environment, Universidad de Alcalá, Calle Colegios 2, 28801 Alcalá de Henares, Spain; emilio.chuvieco@uah.es
[2] Instituto Pirenaico de Ecología, Consejo Superior de Investigaciones Científicas (IPE-CSIC), 50059 Zaragoza, Spain; magi.franquesa@ipe.csic.es
[3] Laboratorio de Climatología y Servicios Climáticos (LCSC), CSIC-Universidad de Zaragoza, 50018 Zaragoza, Spain
[4] Department of Geography and Land Management, University of Zaragoza, 50009 Zaragoza, Spain
[5] GEOFOREST Group University Institute of Research in Environmental Sciences—IUCA, University of Zaragoza, 50009 Zaragoza, Spain
* Correspondence: mclara.ochoa@uah.es

Abstract

A substantial proportion of wildfires in Mediterranean regions continue to be recorded without information about the cause or source of ignition, limiting our ability to understand ignition drivers and design effective prevention strategies. In this study, we develop a spatially harmonised wildfire database for mainland Spain by integrating ignition records from the Spanish General Fire Statistics (EGIF) with fire perimeters generated from satellite images. We then apply a Random Forest classifier to infer ignition causes for events lacking cause attribution. To interpret model behaviour, we use Shapley Additive Explanation (SHAP) values at both global and local scales. Results indicate that human-caused ignitions are dominant, with intentional and negligence-related fires accounting for 52.13% of all known events, although they are associated with contrasting climatic and land-use settings. Negligence-related fires tend to occur under hot, dry and windy conditions, often in agricultural interfaces, whereas intentional fires are more frequent under cooler and wetter conditions and in areas with higher population density and land-use change. Lightning-caused fires represent a small fraction of total ignitions (3%) but exhibit a distinct climatic signature, occurring primarily in sparsely populated areas, under intermediate moisture conditions, and often leading to larger burned areas. Despite strong overall model performance (F1-score = 0.82), minority classes (e.g., lightning and fire rekindling, 0.17%) remain challenging to classify, reflecting both data imbalance and uncertainty in causal attribution. Overall, the combined use of machine learning and explainable AI provides a coherent spatial characterisation of wildfire ignition drivers across mainland Spain, highlights systematic differences among ignition causes, and identifies key limitations in existing fire cause records. This framework represents a practical step towards improving fire cause information by integrating remote sensing products with field-based fire reports, thereby supporting more targeted and evidence-based fire risk management.

Keywords: wildfire dataset; ignitions causes; Spain; random forest; SHAP

1. Introduction

Wildfires represent one of the most pressing environmental challenges in Mediterranean regions, yet the causes behind many of these events remain uncertain. Globally, a

considerable proportion of recorded wildfires continue to be classified as having 'unknown' or 'undetermined' origins, revealing a significant gap in information on fire occurrence [1]. Across Europe, about 50% of all wildfires documented between 1999 and 2016 were reported with unknown causes [2], while in Spain between 1991 and 2000, around 21% of fire events lacked an identified ignition source (Spanish General Fire Statistics, 2002). This uncertainty is central to contemporary wildfire management because incomplete or inaccurate attribution limits the ability of researchers and practitioners to understand ignition drivers, model fire occurrence, and design effective prevention strategies. According to previous studies [1,3–5], the relevance of investigating unknown causes can be summarised with three main aspects: the magnitude of the missing data problem, its implications for scientific modelling, and its critical role in fire prevention and mitigation.

Understanding how wildfires vary according to their ignition source—whether natural or human-caused, and within the latter, accidental or intentional—is essential for designing effective prevention policies. Although less frequent in number, natural ignitions can produce some of the largest and most severe fires in Mediterranean regions, particularly when they occur in remote, forested, and mountainous areas where human presence is limited. Their ignition efficiency depends strongly on meteorological conditions that influence both lightning flashes and fuel moisture content [6]. Their impact highlights the need for advanced strike monitoring and targeted fuel management in wilderness zones [3]. Human-caused ignitions dominate the contemporary fire regime in southern Europe, accounting for over 95% of recorded fires [1]. Intentional ignitions are often associated with socio-economic factors, land-use conflicts, or opposition to conservation measures; in Spain, they tend to concentrate in autumn and winter near the wildland–urban and agricultural interfaces [3]. Accidental ignitions—linked to agricultural burning, machinery, and power line sparks, or other human-related activities—follow marked seasonal patterns and commonly occur near settlements and transport corridors [3]. By extending the ignition season beyond the peak fire weather period (e.g., agricultural burning in early spring), human activities widen the temporal window of risk while increasing the complexity of prevention.

Machine learning (ML) has become a cornerstone for advancing the understanding, modelling, and management of wildfires, offering strong capabilities for handling the complexity and nonlinearity of large environmental datasets [7]. Compared to traditional statistical methods, ML algorithms such as Random Forest (RF) have shown higher predictive accuracy in modelling human-caused fire occurrences than standard regression analysis, achieving AUC values of 0.75 and 0.73 in studies conducted in Spain [8]. These approaches also allow researchers to assess the relative importance of climatic, vegetation, and anthropogenic variables, improving the diagnostic understanding of ignition dynamics and spatial variability [3]. Beyond prediction, ML has been instrumental in addressing the long-standing issue of fires with unknown causes, enabling the classification of ignition sources with notable accuracy [2]. Such applications highlight the potential of ML not only to predict fire occurrence but also to uncover the underlying drivers of ignition patterns, while explainable artificial intelligence (XAI) techniques such as Shapley Additive Explanations (SHAP) improve interpretability and stakeholder confidence by revealing how specific variables influence model outputs [2,9,10].

The increasing use of ML in wildfire science and management has therefore highlighted the need for XAI techniques, which address both the complexity of ML models and the relevance of decisions derived from them [2,10,11]. Many algorithms, while highly effective at detecting complex patterns in large datasets, are often considered "black boxes" because the internal decision-making processes can be opaque [10]. Techniques such as SHAP enable researchers to evaluate the primary drivers of wildfire ignitions across terri-

https://doi.org/10.3390/fire9040138

tories to quantify the influence of specific variables on global and local predictions, and improve the credibility of model outputs for prevention planning [11]. Importantly, XAI facilitates the design of targeted prevention strategies by revealing not only where a fire is likely to occur but also why, allowing interventions to focus on key environmental and anthropogenic factors.

This study pursues three main objectives. First, it aims to harmonise the ESFire30 burned area perimeters derived from remote sensing with the EGIF database, which contains field-based information on wildfire causes. By spatially matching fire perimeters from ESFire30 with ignition records from EGIF, this objective seeks to assign observed fire causes from EGIF to the corresponding burned area perimeters in ESFire30. Second, the study aims to infer the probable causes of wildfires for which no causal information is available by using machine learning methods. This objective addresses the high proportion of fires with unknown causes and responds to the critical importance of fire causality information for wildfire management. Third, the study aims to apply explainable artificial intelligence (XAI) techniques to identify the environmental and anthropogenic variables that most strongly contribute to a better understanding of cause-specific fire patterns and help identify areas that are more vulnerable to different types of wildfire ignitions.

2. Methodology

2.1. Study Area

The study area comprises mainland Spain, which covers approximately 493,000 km^2, excluding the Balearic and Canary Islands. The region is characterised by a complex topography that includes extensive plateaus (i.e., the Meseta Central, including the northern and southern subplateaus), several major mountain ranges such as the Pyrenees, the Cantabrian Mountains, the Iberian System, and Sierra Nevada, as well as a diverse set of coastal plains. Peninsular Spain experiences a predominantly Mediterranean climate, marked by hot, dry summers and mild, wet winters, although substantial spatial variability exists. The northern Atlantic regions (e.g., Galicia, Asturias, Cantabria and the Basque Country) have a temperate oceanic climate with higher rainfall, while the southern and eastern areas (e.g., Andalusia, Murcia, and Valencia) experience semi-arid conditions. This climatic and topographic diversity supports a wide range of ecosystems and land-use types, from Mediterranean shrublands and pine forests to dense broadleaved forests, and from rainfed agricultural areas in the hinterlands to densely populated coastal zones.

2.2. Input Fire Datasets

Two databases were used to conduct this study, each providing complementary information on wildfire occurrence. The ESFire30 contains annual fire perimeter data for mainland Spain from 1985 to 2023, including all burned areas larger than 5 ha. It was derived from Landsat satellite imagery using a semi-automatic processing algorithm, with post-processing refinements (Ochoa et al., 2026, under review). In total, ESFire30 contains 119,498 fire perimeters, accounting for 5,977,591 ha of burned area [12].

Ignition information was obtained from the official Spanish fire database, designated as EGIF (Estadística General de Incendios Forestales: Spanish General Fire Statistics). This database is maintained by the Ministry for the Ecological Transition and the Demographic Challenge (MITECO). The database integrates data reported by regional forest and emergency services, ensuring national consistency through a unified reporting protocol. The EGIF consists of a comprehensive and standardised record of wildfire events across Spain since the 1960s, with a systematic national coverage beginning in 1968. It compiles detailed information for each fire event, including the date, ignition cause, size, land cover affected, suppression resources, and environmental conditions at the time of ignition. The database

https://doi.org/10.3390/fire9040138

is currently available in Excel format. Since 2012, fire event records include spatial coordinates for ignition locations; however, for earlier years—and for a substantial portion of the dataset—these coordinates are not available. Consequently, the primary spatial reference system used in EGIF is a 10 × 10 km grid.

Fire causes are categorised into several groups, including lightning, negligence or accident, intentional fires, fire rekindling, and unknown causes. Due to its temporal coverage, the EGIF is one of the most valuable sources for analysing wildfire patterns, drivers, and trends in Spain. The dataset is regularly updated and publicly available at https://servicio.mapa.gob.es/incendios/Search/Publico, last accessed on 1 January 2026. While data are accessible up to 2023 at the provincial level, the most comprehensively verified national dataset dates to 2016.

2.3. Spatialisation of EGIF Ignition Records and Attribution of Fire Cause Information to ESFire30 Perimeters

Although highly valuable, the EGIF database presents important spatial limitations. For fires recorded before 2012, geographic coordinates—representing the reported location of ignition—are often missing or, when included, frequently contain positional errors, likely introduced during data transfer to the database. Prior to 2012, fire locations were referenced using an outdated sheet and grid system (sheets of 110 × 150 km and grids of 10 × 10 km) derived from the 1:250,000 military map, which is still used for official fire reporting (Figure 1a). To associate EGIF ignition information with ESFire30 fire perimeters, we applied a methodology based on Rodrigues and Riva [8] to spatialise the ignition records. Based on the sheet and grid system (Figure 1a), we first created a 'burnable area' layer by masking urban areas, water bodies, and a 100 m buffer around them to avoid placing ignitions on fully non-burnable surfaces. Random points were then generated within each grid cell, with the number of points matching the number of recorded EGIF observations for the corresponding sheet and grid (Figure 1b).

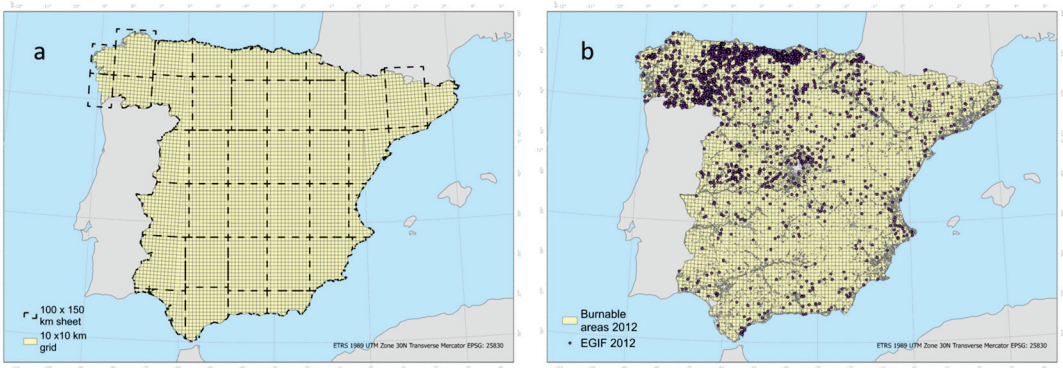


Figure 1. Spatialised ignition points from EGIF database: (**a**) example of the sheet grid system used to locate ignitions; (**b**) resulting spatialised ignition points over burnable areas.

Once the EGIF ignition points were generated, they were integrated into the ESFire30 dataset. However, there were various inconsistencies between the two datasets. In many grid cells, the number of fire perimeters and ignition point records did not coincide. Some cells contained more perimeters than ignition points, others had more ignition points than perimeters, and in some cases only one of the two datasets contained records. To assign a cause to each fire perimeter, we matched EGIF ignition points and ESFire30 fire perimeters within the same grid cell when their burned areas were similar (±10 ha). When such a match was found, the cause from the source ignition point was assigned to the corresponding polygon. If no ignition point with a comparable burned area was available, the predominant cause recorded in that grid cell was used instead, reflecting the

https://doi.org/10.3390/fire9040138

strong spatial clustering and local predominance typically observed in fire cause patterns (MITECO, 2019). To evaluate the reliability of this approach, we calculated the proportion of fires in each grid cell corresponding to the assignment majority cause, confirming that, in most cells, the predominant cause represents the majority of events. Finally, if a grid cell contained ESFire30 polygons but no EGIF ignition points, the polygon was classified as having an unknown cause. The database records the origin of each assignment in a field labelled 'Method', where a value of 1 indicates a similarity-based burned area match and a value of 2 indicates a majority-based assignment. Before modelling, both spatial and comparative analyses were carried out to evaluate the consistency of the integrated dataset.

2.4. Model

In this study, we developed a model to infer wildfire causes for those fire perimeters that did not have an associated cause in the EGIF database. In this way, we should be able to better understand fire drivers, and provide additional insights to improve fire prevention. We used a Random Forest classifier designed to infer the causes of fire events with unknown origin and to identify the environmental and demographic variables that best explained these patterns.

Prior to modelling, spatial analyses were conducted to identify where wildfires were most prevalent according to their cause, and to calculate the proportion of each cause category to establish an expected baseline distribution. These exploratory steps were essential for understanding the underlying spatial and statistical behaviour of the causes in our dataset, ensuring that the results obtained after the modelling process were coherent and meaningful. By characterising these patterns in advance, we were able to better interpret the model outputs and verify that they aligned with the observed dynamics in the study area.

Before beginning the classification stage, it was necessary to address the class imbalance in the target variable. Some cause categories were clear minorities, which can lead to biased model predictions. To alleviate this issue, we applied the Synthetic Minority Oversampling Technique (SMOTE), a widely used method in ecological and environmental modelling to handle class imbalance [2,11,13]. SMOTE generates synthetic samples of the minority classes rather than simply duplicating existing observations. It does so by selecting a minority class instance and creating new synthetic samples along the line segments connecting that instance with its k-nearest minority neighbours. This approach increases the number of minority samples in a more realistic and informative way, helping the model learn more accurate decision boundaries between classes.

In our case, the dataset was split into training (70%) and testing (30%) sets implementing a stratification procedure based on the original class proportions. SMOTE was applied only to the training set (k_neighbours = 2), ensuring that the test data remained independent for validation. A small number of neighbours was selected to ensure that the synthetic samples remained close to the original minority class observations, preserving the local structure of the feature space. SMOTE resampling improves the classifier's ability to correctly identify minority class events. While applying SMOTE in our model does not significantly boost the overall F1-score (82%), it does enhance the performance metrics for individual classes. Importantly, when the trained classifier is applied to the unseen test set, it performs inference in the machine learning sense: predicting the likely ignition cause for new wildfire instances based on parameters learned during training. This inference should not be interpreted as established causal relationships, but rather as generating predictions for data not used during model fitting.

To optimise model performance and prevent overfitting, cross-validation techniques were applied to determine the optimal set of hyperparameters. A grid search proce-

https://doi.org/10.3390/fire9040138

dure (GridSearchCV from the scikit-learn library [14]) was implemented using a five-fold cross-validation scheme, testing combinations of parameters such as the number of trees (n_estimators), the maximum tree depth (max_depth), the minimum number of samples required for node splitting (min_sample_split), the minimum number of samples per leaf (min_sample_leaf), and the maximum number of features considered at each split (max_features). The model achieved a mean cross-validated accuracy of 0.79 ± 0.004 (95% CI: 0.788–0.792), indicating stable performance across repeated folds. The final model configuration corresponded to the combination of parameters that achieved the highest mean cross-validation accuracy.

Once the classes were balanced, we trained an RF classifier to model wildfire causes. The RF algorithm is an ensemble learning method that builds multiple decision trees during training and combines their outputs to produce more accurate and robust predictions. The model was configured with the following parameters:

- n_estimators = 200, ensuring a sufficiently large number of trees to reduce variance.
- max_depth = None, allowing trees to grow fully and capture complex patterns.
- max_features = 'sqrt', selecting a subset of variables at each split to increase model diversity and reduce overfitting.
- min_samples_split = 5 and min_samples_leaf = 1, controlling how trees split and ensuring sufficient flexibility to learn from the data.
- random_state = 42, ensuring reproducibility.
- class_weight = 'balanced', giving more weight to minority classes to reduce bias toward majority categories.

To interpret the model and better understand the factors related to wildfire causes, we applied SHAP. This approach quantifies the contribution of each predictor variable to individual predictions, providing a transparent and interpretable way to identify which variables exert the strongest influence on each fire cause classification.

In addition to this global interpretation, we performed a local spatial analysis to determine which variables exert the greatest influence in specific geographic regions. This approach makes it possible to detect how the importance of individual drivers varied across the landscape, highlighting areas where specific climatic or environmental factors played a more dominant role.

2.4.1. Dependent Variable

For this model, the dependent variable is the ignition cause. The 119,498 fires recorded in the ESFire30 dataset were classified into five categories as reported in the EGIF (see Table 1):

Table 1. Dependent variable metrics.

Type	Class in ESFire30	EGIF Code	Total Number of Fires	Total Burned Area (ha)	Maximum Patch Size (ha)	Percentile 90 (ha)	Proportion (%)
Lightning	1	100	2966	309,417	19,978	85	3
Negligence	2	210 to 399	11,809	563,429	25,675	58	10
Intentional	3	400	50,344	2,021,564	14,060	68	42.13
Unknown	4	500	54,167	2,228,235	28,765	64	45.32
Fire rekindling	5	600	212	21,341	11,773	74	0.17

2.4.2. Independent Variables

To identify relevant human-related covariates, we drew on established methodologies and findings from prior research [5,8,15–17]. We also considered regional- and European-scale studies [11,18–20], as well as insights from a review on fire occurrence modelling [1].

https://doi.org/10.3390/fire9040138

The selected variables represent widely recognised indicators of fire occurrence and follow the typology of factors and drivers described by [21,22]. Some variables required transformations prior to integration into the model, whereas others were used in their original form. Table 2 summarises all selected variables. The transformation procedures are described below:

- Road density: To assess spatial accessibility in relation to wildfires, we generated secondary road density maps using a kernel density approach applied to linear vector data. Secondary roads were selected because they are closely associated with fire suppression access and human presence, both of which influence ignition risk [1]. The road network was converted into regularly spaced points (every 10 m) and rasterised at 30 m resolution, with each cell representing road density values. This continuous variable serves as a proxy for spatial accessibility: lower values correspond to remote or difficult to access areas, while higher values indicate easier access and greater human presence. Because historical road data are limited, we used a single road layer (https://centrodedescargas.cnig.es/CentroDescargas/redes-transporte, last accessed on 1 January 2026) supported by evidence that rural and secondary road networks in Spain have changed little in the past three decades [23]. For each burned area perimeter, we extracted the mean road density value to ensure that the metric reflects average accessibility conditions and is not biased by polygon size.
- Interfaces: Interfaces are spatial zones representing areas of contact or interaction between forest or wildland vegetation and other land-use categories. They delineate the proportion of the landscape where human activities or different land-cover types meet flammable natural vegetation, creating conditions that may influence both the likelihood of ignition and the behaviour of wildfires.

In this study, interfaces were derived from the SIOSE (https://centrodedescargas.cnig.es/CentroDescargas/siose, last accessed on 1 January 2026) and Corine Land Cover (https://centrodedescargas.cnig.es/CentroDescargas/corine-land-cover, last accessed on 1 January 2026) datasets for their corresponding years. A 250 m buffer was applied around forest and wildland classes and overlapping areas with the corresponding land-use class of interest (e.g., urban and agricultural) were identified. The 250 m buffer distance was selected to capture the immediate spatial context surrounding for each class while maintaining consistency with the spatial resolution of the burned area dataset (30 m). Similar distances have been used in studies analysing wildland–urban interfaces, as they represent the local zone where interactions between vegetation and anthropogenic land uses can influence ignition probability and fire dynamics [24,25]. The resulting variables quantify the total interface area (ha) within each burned area perimeter, representing the cumulative area of contact that burned. Specifically, five interface variables were generated to characterise different types of land-cover contact: the forest–agricultural interface (FAI), wildland–urban interface (WUI), wildland–agricultural interface (WAI), wildland–forest interface (WFI), and forest–urban Interface (FUI). This approach produces continuous variables, which improve the performance of RF models in representing gradients of human presence, land-use interaction, and potential ignition pressure. In general, interfaces capture the spatial coupling between vegetation and human-related land covers, functioning as indicators of socio-ecological processes.

- Rural areas (1985–1999 and 2000–2021): From 1985 to 1999, rural areas were derived from the Global Human Settlement Population Layer (GHS: https://human-settlement.emergency.copernicus.eu/ghs_pop.php, last accessed on 1 January 2026). The 1985 layer was assigned to 1985–1989, the 1990 layer to 1990–1995, and the 1995 layer to 1995–1999, following the five-year structure of the dataset [26]. The GHS dataset, developed by the Copernicus Emergency Management Service, provides

https://doi.org/10.3390/fire9040138

globally consistent gridded estimates of residential population density by combining census data with satellite-derived built-up area information. From 2000 to 2021 rural areas were obtained using the WorldPop unconstrained population density dataset at a 1 km spatial resolution (https://www.worldpop.org/, last accessed on 1 January 2026) [27]. WorldPop redistributes census counts using an RF-based dasymetric modelling approach to generate high-precision gridded population estimates. For 2021, the layer 2020 was used.

Following Law 45/2007 on Sustainable Rural Development, grid cells were classified as rural when population density ranged between 1 and 100 inhabitants per km^2, consistent with the definition of rural municipalities <30,000 inhabitants and <100 inhabitants/km^2 [28]. This method delineates sparsely populated areas based on actual population distribution rather than administrative boundaries, thereby improving spatial accuracy. Urban areas were assigned a value of zero—not to indicate absence of population, but to represent absence of rural population specifically. For each burned area perimeter, rural population values were summed across all intersecting pixels.

- Land-use change: Land-use change was derived from the Land Use, Land-Use Change and Forestry (LULUCF) database: https://www.eea.europa.eu/en/analysis/indicators/greenhouse-gas-emissions-from-land, (last accessed on 1 January 2026) which spans 1970–2021 [29]. We extracted pixels representing land abandonment, defined as land-cover transitions to wildlands or forests. Since LULUCF data are organised in multi-year intervals, each burned area perimeter was assigned to the interval containing its fire year to ensure temporal consistency. The total area of land-use change within each burned area perimeter was then calculated.
- Grazing: To account for livestock pressure, we employed the 2015 global livestock-density layer expressed in heads/km^2 (https://www.fao.org/land-water/land/land-governance/land-resources-planning-toolbox/category/details/en/c/1236449/, last accessed on 1 January 2026). Although this dataset represents a single year, it provides a suitable proxy for the entire study period given the lack of comparable high-resolution historical spatial datasets. As highlighted by Serrano-Zulueta, Gómez-Sal [30], traditional pastoral systems in Spain have undergone a sustained decline since the mid-20th century, driven by rural depopulation, structural shifts in livestock production, and policies favouring intensification. Similarly, Acebes, Iglesias-González [31] emphasise that grazing systems—particularly extensive and transhumant practices—are currently facing collapse in many regions. Given this long-term downward trend, using the 2015 livestock density layer likely represents an upper bound or stabilised snapshot of grazing pressure, offering a conservative and consistent estimate across the study period. For each burned area perimeter, we extracted the mean livestock density value.
- Climatic variables: Climatic variables were obtained from gridded meteorological datasets provided by the Spanish Meteorological Agency (AEMET) and other operational sources used in the Spanish meteorological drought monitoring system. This system is based on quality-controlled station observation and spatial interpolation at approximately 1.1 km resolution and provides long-term climatic information for Spain (https://monitordesequia.csic.es/historico, last accessed on 1 January 2026) [32]. The variables used in this study were potential evapotranspiration (eto), relative humidity (hr), incoming solar radiation (in_), precipitation (pr), dew point temperature (td), maximum temperature (tmax), minimum temperature (tmin), vapour pressure deficit (vpd), and wind speed (w). For each burned area perimeter, summer mean values (June, July, and August) were calculated for each climatic variable.

https://doi.org/10.3390/fire9040138

- Standardised Precipitation–Evapotranspiration Index (SPEI): To characterise summer climatic wetness and dryness in the months preceding each fire event, we computed the 3-month Standardised Precipitation–Evapotranspiration Index (SPEI-3). SPEI-3 was calculated for each fire perimeter using the climatic variables described above (precipitation and reference evapotranspiration), following the standard formulation based on the monthly climatic water balance and fitted to a log-logistic distribution as proposed by Vicente-Serrano, Beguería [33]. For each fire perimeter and year from 1985 to 2021, we extracted the August SPEI-3 value, reflecting drought conditions accumulated over June, July, and August—the main period of fuel drying in Mediterranean climates. This metric captures seasonal moisture anomalies relevant to fire occurrence and provides a standardised measure of short-term climatic dryness or wetness prior to ignition.
- Normalised Difference Vegetation Index (NDVI): Seasonal vegetation conditions were derived from the VIIB NDVI database, which provides biweekly observation at a 1.1 km resolution for Peninsular Spain from 1981 to the present [34]. For each burned area perimeter, NDVI values were first extracted for all 1044 biweekly dates available between July 1981 and December 2024 using area-weighted averaging, producing a continuous NDVI time series representative of mean vegetation conditions within each polygon. These biweekly values were then aggregated by calendar year to obtain seasonal means for spring (MAM: March–May) and summer (JJA: June–August), typically comprising six observations per season.

To account for interannual variability relative to historical norms, the resulting seasonal NDVI series were standardised following procedures commonly used for indices such as SPI and SPEI. For each perimeter and year, the variable included in the model was ndvi_mam, capturing vegetation conditions during the key growing season preceding and shaping summer fuel flammability.

Table 2. Biophysical and human explanatory variables.

Variable	Source	Temporal Extent	Resolution
Population density 1985–1999	[26]	1985; 1990; 1995	1 km
Population density 2000–2021	[27,35]	2000–2020	1 km
Secondary Roads density	[36]	2020–2025	Minimum mappable unit 5 m
Forest–agricultural interface (FAI)	[37,38]	SIOSE: 2005, 2009, 2011, 2014. CORINE: 1990, 2000, 2006, 2012, 2018	SIOSE: Minimum mappable unit 1 ha CORINE: Minimum mappable unit 25 ha
Wildland–urban interface (WUI)	[37,38]	SIOSE: 2005, 2009, 2011, 2014. CORINE: 1990, 2000, 2006, 2012, 2018	SIOSE: Minimum mappable unit 1 ha CORINE: Minimum mappable unit 25 ha
Wildland–agricultural interface (WAI)	[37,38]	SIOSE: 2005, 2009, 2011, 2014. CORINE: 1990, 2000, 2006, 2012, 2018	SIOSE: Minimum mappable unit 1 ha CORINE: Minimum mappable unit 25 ha
Wildland–forest interface (WFI)	[37,38]	SIOSE: 2005, 2009, 2011, 2014. CORINE: 1990, 2000, 2006, 2012, 2018	SIOSE: Minimum mappable unit 1 ha CORINE: Minimum mappable unit 25 ha

https://doi.org/10.3390/fire9040138

Table 2. *Cont.*

Variable	Source	Temporal Extent	Resolution
Forest–urban interface (FUI)	[37,38]	SIOSE: 2005, 2009, 2011, 2014. CORINE: 1990, 2000, 2006, 2012, 2018	SIOSE: Minimum mappable unit 1 ha CORINE: Minimum mappable unit 25 ha
Rural areas 2000–2021	[27,35]	2000–2020	1 km
Rural areas 1985–1999	[26]	1985; 1990; 1995	1 km
Land change	[29]	1985–2021	25 m
Grazing (goats, sheep and cattle /km^2)	[39]	2015	1 km
NDVI March, April, and May (MAM)	[40]	1985–2021	1 km
Climatic variables Jun, July, and August (JJA) eto: evapotranspiration (potential) hr: relative humidity in_: incoming solar radiation pr: precipitation td: dew point temperature tmax: maximum temperature tmin: minimum temperature vpd: vapor pressure deficit w: wind speed	[32]	1985–2021	1 km
Spei 3_august and (Evapotranspiration Standardised Precipitation)	computed as in [33] using de SPEI R package	1985–2021	1 km

A correlation matrix was then calculated for all predictor variables to determine whether any were collinear and thus providing redundant information. The analysis revealed clear patterns of association among climatic variables, given that related atmospheric and environmental processes influence many of them. As a result, variables representing different aspects of the climate system tended to vary together in a consistent manner. Importantly, even if some variables were correlated, retaining them in the model was still valuable for classifying wildfires causes, as each variable may capture unique aspects of environmental conditions and contribute to improving the model's accuracy and robustness.

3. Results

3.1. Spatial Distribution and Proportion of Wildfires by Causes

Figure 2a (EGIF) and 2b (ESFire30) show that the proportion of fires attributed to natural causes and fire rekindling remains relatively stable throughout the study period. However, several notable differences emerge between the two datasets. In ESFire30 (Figure 2b), the category of unknown causes dominates in nearly all years, followed by intentional and negligence-related fires, which show moderate fluctuations over time. In contrast, EGIF (Figure 2a) displays a lower proportion of unknown causes and a higher relative share of intentional fires during most of the period. Negligence-related fires appear comparatively more stable and less variable in EGIF than in ESFire30. Interannual variability is also more pronounced in the ESFire30 series, particularly for the intentional and unknown categories, whereas EGIF exhibits smoother year-to-year changes.

Figure 2c shows the distribution of fire causes in ESFire30 after the classification. The proportion of the different cause categories become more similar to those observed in EGIF. This is especially visible in 2020 and 2021. In these years, ESFire30 originally had a very large share of fires recorded as "unknown", mainly due to missing information. The

https://doi.org/10.3390/fire9040138

classification led to a distribution of fire causes in ESFire30 that was closer to the baseline established by EGIF.

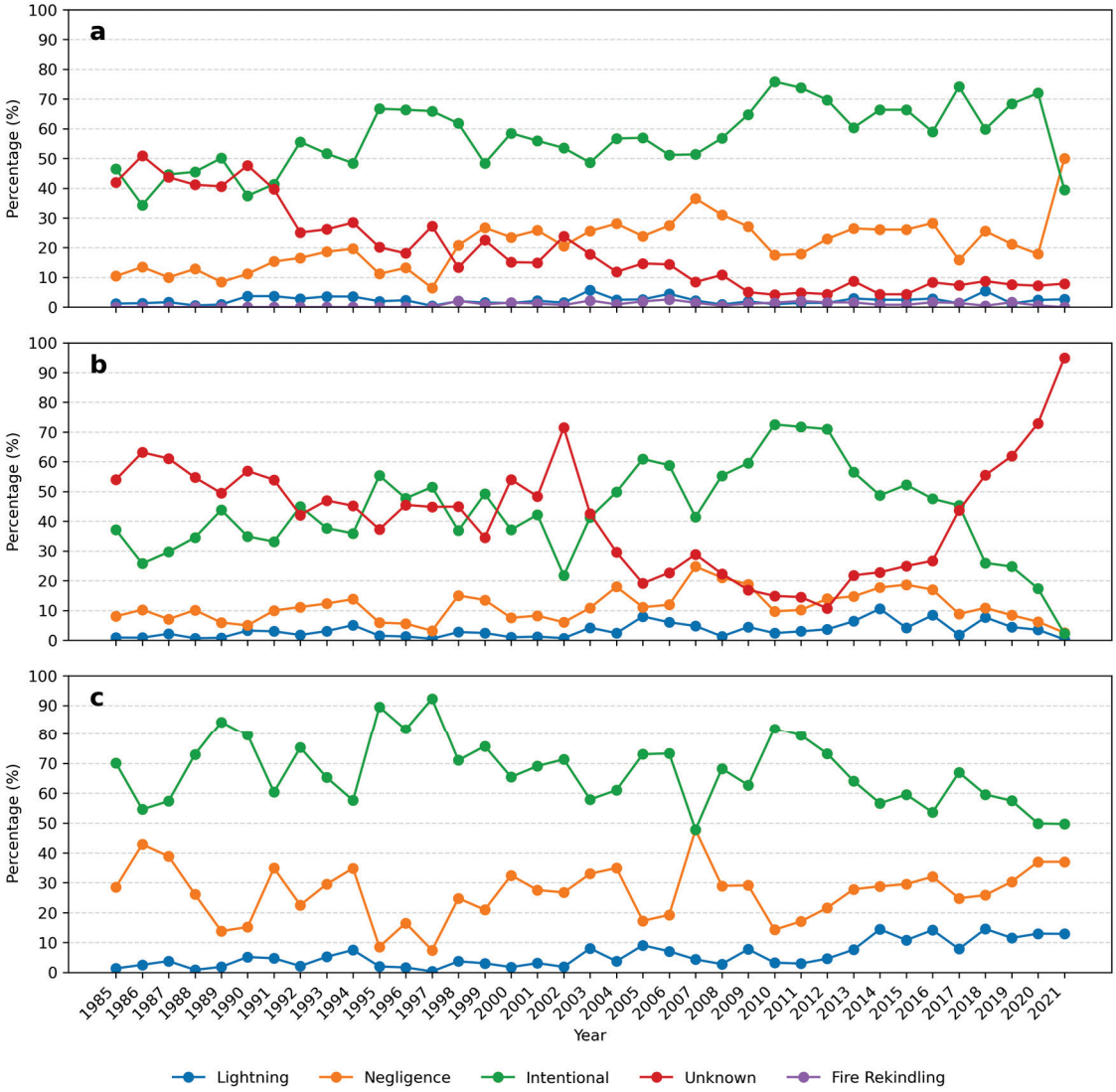


Figure 2. Annual percentage of wildfires by cause from 1985 to 2021: (**a**) shows EGIF data, (**b**) shows ESFire30 data, and (**c**) shows ESFire30 after classification.

The spatial distribution of wildfire causes reveals clear and contrasting patterns across Peninsular Spain (Figure 3). The maps display the density of wildfires by cause at a 30 m resolution, corresponding to the native resolution of the ESFire30 dataset (Ochoa et al. 2026: under review). The delineated regions correspond to the pyro-regions defined by Trigo, Sousa [41], who applied a cluster analysis to normalised burned area data from 66 administrative units across mainland Spain to identify regions with similar fire regimes. This analysis yielded four ecologically coherent clusters—northwest (NW), north (N), southwest (SW), and east (E)—which reflect distinct patterns in topography, vegetation cover and climatic conditions.

Human-caused wildfires (Figure 3a) are predominantly concentrated in the NW pyro-region, particularly along the border with Portugal. A notable density is also observed in the N pyro-region, with lower concentrations in the SW and E regions. In contrast, natural cause wildfires are primarily located within the eastern pyro-region. Figure 3b shows that wildfires of natural origin tend to cluster in the E region. Figure 3c displays wildfires of

https://doi.org/10.3390/fire9040138

unknown cause, which are widely distributed across the territory, with particularly high densities in the NW and N pyro-regions and smaller clusters in the SW and E regions.

Given the high proportion of wildfires with unidentified causes, we completed the attribution of causes using the RF model that was previously described, inferring the most likely causes of each fire perimeter from the auxiliary variables of the model.

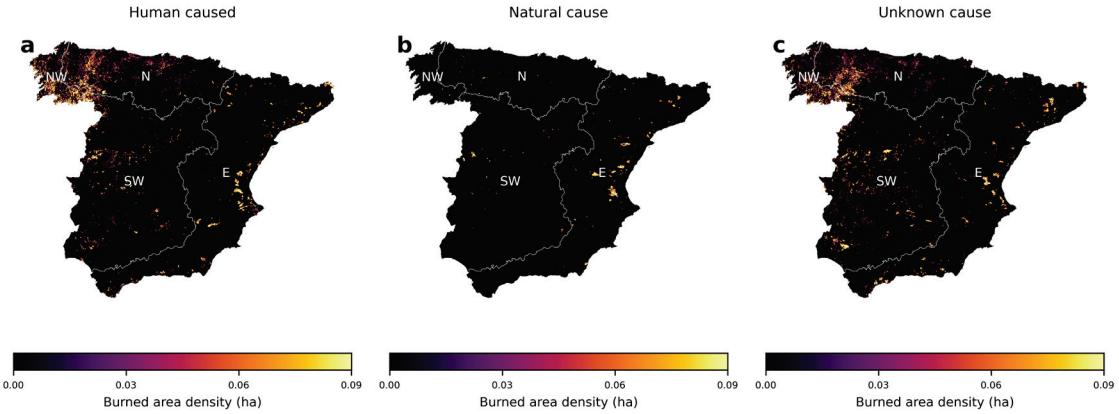
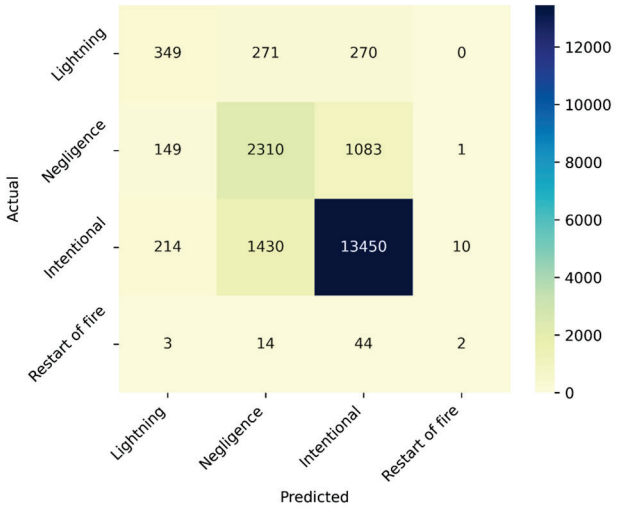


Figure 3. Wildfire causes from 1985 to 2021 as presented by ESFire30.

3.2. Model Performance

The performance metrics of the RF estimations (Figure 4) summarises the classification results across the four fire cause categories. The confusion matrix shows that the intentional class has the highest number of correctly predicted cases, with relatively few misclassifications. In contrast, the negligence and lightning classes exhibit greater dispersion, with a noticeable number of instances misclassified as intentional. The fire rekindling category contains very few observations overall and shows substantial overlap with other categories, indicating limited correct identification.



Class	Precision	Recall	F1-Score
Lightning	0.49	0.39	0.43
Negligence	0.57	0.65	0.61
Intentional	0.91	0.89	0.90
Restart of fire	0.15	0.03	0.05
Accuracy global			0.82
F1-score global			0.82
Recall global			0.82
Precision global			0.82

Figure 4. Confusion matrix and performance parameters per class.

The intentional class achieves the highest values across all metrics, with a precision of 0.91, recall of 0.89 and F1-score of 0.90. Negligence shows consistent values (precision = 0.57, recall = 0.65, and F1-score = 0.61), while lightning records lower but still consistent values (precision = 0.49, recall = 0.39, and F1-score = 0.43). Fire rekindling presents very low precision (0.15) and recall (0.03), resulting in a minimal F1-score (0.05). Overall, the model achieves a global accuracy and F1-score of 0.82, indicating strong general

https://doi.org/10.3390/fire9040138

predictive performance. However, performance varies notably among classes, with the model showing great difficulty in identifying minority classes despite the application of class balancing techniques.

3.3. Variable Importance

Based on the SHAP analysis, distinct environmental and anthropogenic patterns were identified for each fire cause (Figure 5). High evapotranspiration, high precipitation, and large burned areas, combined with low NDVI, low population density, and low grazing activity, characterised lightning-caused fires. These fires were also associated with medium maximum temperatures, medium wind speeds, a strong presence of the WFI, greater distances to roads, and minimal land-use change.

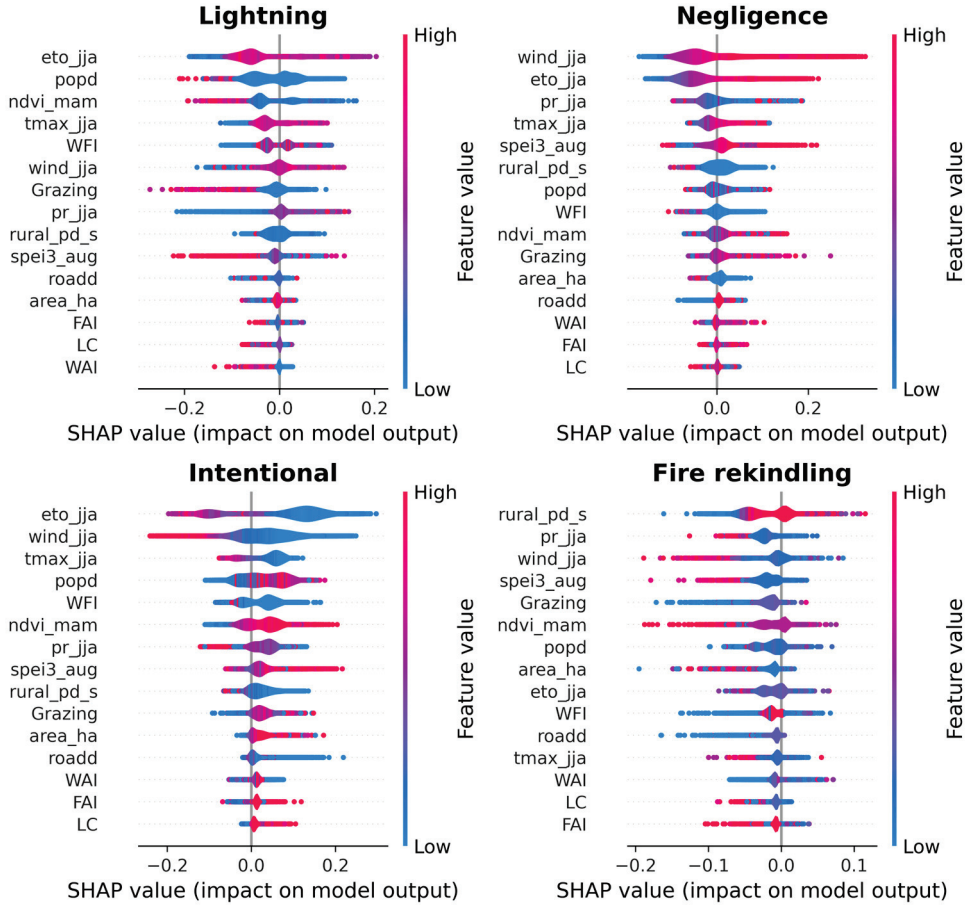


Figure 5. SHAP summary plots showing the importance and effect of predictor variables for each ignition cause. The colour of each dot represents the feature value and its position along the row represents the SHAP value for a given feature and instance. Positive values (right of the centred line) indicate a higher probability of belonging to a specific class, while negative values (left of the centred line) indicate a lower probability of belonging to a specific class.

Negligence-related fires exhibited a markedly different profile, with high wind speeds, high evapotranspiration, high maximum temperatures, and high SPEI values, coupled with low precipitation, low rural population, and low population density. These events showed high NDVI and grazing values, greater distances to roads, and were linked to high WAI and FAI, but low land-use change.

Intentional fires contrasted sharply with negligence-related ones. They occurred under low evapotranspiration, low wind speeds, and low maximum temperatures, yet in areas with high population density and high NDVI. These fires were associated with high

https://doi.org/10.3390/fire9040138

precipitation and large burned areas, high FAI, low WAI, and high land-change values, but with low rural population, high grazing activity, and high SPEI.

Finally, fire rekindling events were characterised by high rural population and high grazing activity, but low precipitation, low wind speeds, low evapotranspiration, and low SPEI. These events generally occurred in areas with medium NDVI, low population density, and low values across all interface and land-change metrics.

Overall, the most notable contrast emerged between intentional and negligence-related fires: while both are linked to human-related factors, negligence fires occurred under hotter, drier, and windier conditions with high evapotranspiration and high SPEI, whereas intentional fires were associated with cooler, wetter environments and denser populations. This distinction highlights different underlying mechanisms and contextual drivers behind these two anthropogenic fire causes.

An additional observation concerns the role of the burned area across causes. Lightning and intentional fires were associated with large burned areas, indicating that these causes tend to result in more extensive events. In contrast, negligence and fire rekindling showed comparatively low burned area values.

3.4. SHAP's Local Variable Importance

In the eastern regions, lighting-caused fires were primarily influenced by spring NDVI, precipitation, and population density (Figure 6a). The southwestern region showed a similar pattern, with NDVI and precipitation emerging as the dominant predictors, indicating that vegetation greenness and moisture availability are key drivers of natural fire occurrence in this area. In contrast, the N and the NW regions exhibited a more complex structure; NDVI, precipitation, and population density remained important, but additional variables—such as wind speed and maximum temperature—also contributed meaningfully to the predictions. This suggests that a broader combination of environmental factors governs lightning-caused fire occurrence in these regions.

The analysis of intentional fire causes (Figure 6b) across different regions revealed distinct patterns of influence. In the N and NW regions, intentional fires were primarily associated with evapotranspiration, population density, and wind speed, with additional but smaller contributions from land-use variables (interfaces, grazing and land-use change), NDVI and rural population. In the E region, evapotranspiration, population density, land-use factors, and precipitation emerged as the most relevant predictors. In the SW region, evapotranspiration and wind speed were the dominant factors, while population density, land-use variables, and burned area also showed secondary but notable contributions.

Negligence-related fires (Figure 6c) exhibited region-specific patterns of influence. In the central area, where the NW, N, and SW regions intersect, these fires were primarily associated with population density, land-use variables, rural population, and climatic factors. In contrast, across the broader NW, SW, N and eastern regions, evapotranspiration, wind speed and spring NDVI were the most influential variables, emphasising the role of environmental and vegetation conditions in shaping fire occurrence in these areas.

The regional aggregation of SHAP contributions (Figure 6d) reveals that evapotranspiration and wind speed emerge as the dominant predictors across the study area, reflecting their overarching control on fuel moisture and atmospheric conditions. However, notable regional distinctions persist: in the N, NW, and SW regions, population density contributes substantially to the model's output, whereas in the E region precipitation accounts for the largest number of instances. Importantly, this aggregated representation is strongly shaped by the predominance of intentional fire records in the dataset, which drives much of the overall signal.

https://doi.org/10.3390/fire9040138

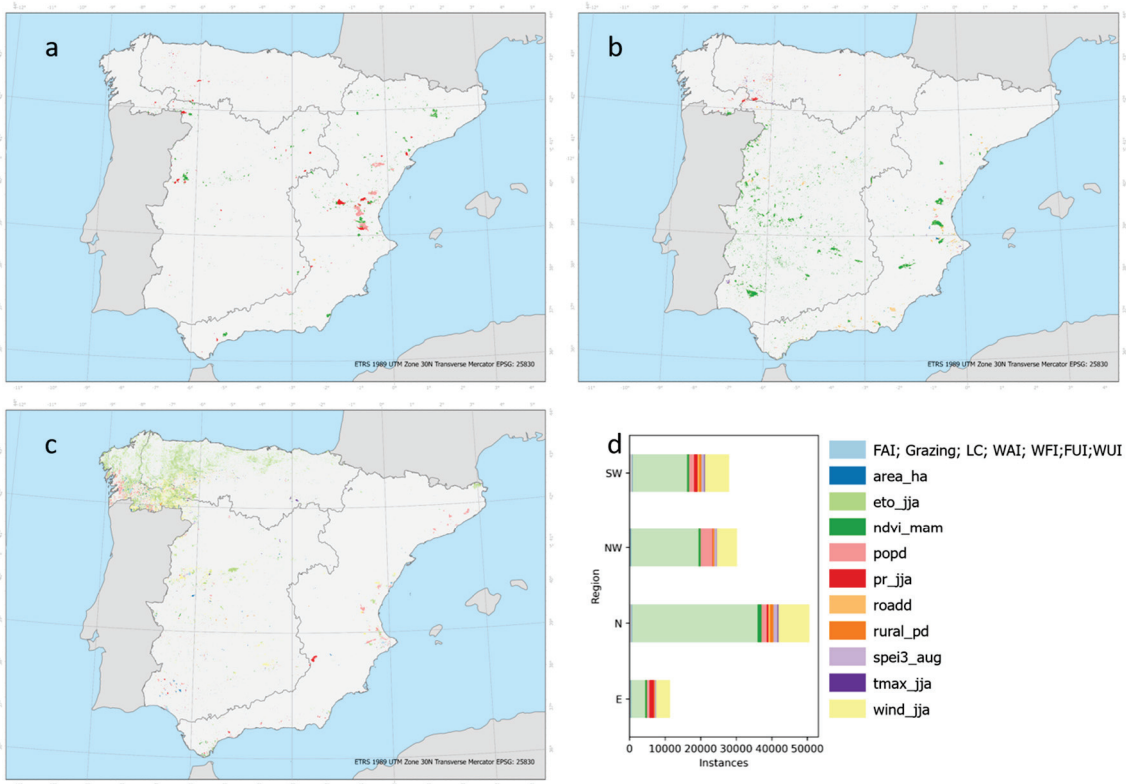


Figure 6. SHAP's local importance for each fire cause category: (**a**) natural, (**b**) negligence, and (**c**) intentional. Panel (**d**) shows the regional aggregated SHAP contribution.

4. Discussion

In this study, we developed a spatially harmonised wildfire database for mainland Spain by integrating ignition records from EGIF with fire perimeters provided by ESFire30, enabling a more consistent attribution of fire causes and a clearer assessment of their spatial distribution. Using this consolidated dataset, we applied an RF classifier to infer the origin of fires lacking an identified cause and to evaluate the environmental and anthropogenic factors most strongly associated with each ignition type. Finally, we employed SHAP-based explainability techniques at both global and local scales to disentangle the differing mechanisms underpinning fire causes, thereby providing an analytical and spatial framework that enhances our understanding of wildfire ignition patterns across the region.

A key methodological distinction concerns how the potential ignition locations of EGIF fires are defined. Previous spatialisation approaches [16] refined ignition areas by intersecting the sheet and grid system within municipal boundaries and forested perimeters, which was appropriate for studies focused specifically on forest-related fire dynamics. In our case, however, representing all fire occurrences is essential, as the complementary dataset used (ESFire30) does not distinguish fires by land-use category. For this reason, our spatialisation procedure avoids restricting ignition locations solely to forested areas and instead retains the full spectrum of burnable environments recorded in EGIF, including those classified as "Other land covers". By masking only clearly non-burnable surfaces—such as urban areas, water bodies, and a surrounding buffer—we ensure that the spatialised EGIF dataset remains compatible with ESFire30 and captures the complete set of fire events relevant to our analysis.

Despite applying SMOTE to balance the training data, the model still learned the dominant classes more effectively, reflecting the underlying class structure resulting from our spatialisation and integration procedures. This is evident in the performance metrics,

https://doi.org/10.3390/fire9040138

where minority categories, such as lightning-caused fires and fire rekindling, show notably lower predictive accuracy. Moreover, the underlying class imbalance can bias the SHAP analysis, as the model assigns stronger explanatory weight to features associated with dominant classes, potentially underrepresenting factors relevant to minor categories. Compared to Bountzouklis, Fox [2]—where lightning ignitions exhibit the highest classification performance—the contrast is striking. In that study, lightning-caused fires were easier to identify and showed very low confusion with other categories, largely because natural ignitions follow a distinct profile that is well captured by reliable field observations. Our results therefore highlight the importance of using input data with higher certainty and better-defined causal labels. Accurate and well-documented field information is essential for training robust models capable of recognising ignition causes effectively, particularly for classes that are naturally less frequent in national datasets. It is worth noting that some categories, such as "unknown" or rekindled fires, are inherently less reliable due to reporting limitations, which can influence both classification performance and interpretability analyses. These uncertainties in label assignment mean that the model's predictions and the subsequent SHAP- based interpretations may be more reliable for well-documented classes than for those with higher reporting uncertainty. Despite these uncertainties, we see that the overall performance of our model is high and even exceeds that reported by Bountzouklis, Fox [2].

For wildfire risk prevention, predicting human-caused fires is challenging due to the complexity of social behaviour, yet research consistently identifies a set of robust predictors that shape their spatial patterns [1]. Accessibility and human presence are among the strongest drivers: proximity to roads, railways, forest tracks, and human–vegetation interfaces (FAI, WUI, and WAI) reliably increases ignition likelihood. Socio-economic conditions such as population density, rural abandonment, unemployment, and ageing population also modulate fire occurrence by shaping how people use and move through the landscape. Climate acts as an enabling factor, with drought and dry weather conditions amplifying the probability that human activities result in ignitions [3,11]. Together, these factors highlight a complex interplay between people, land-use, and environmental conditions, providing essential context for interpreting the patterns observed in our own results.

Building on these established drivers, our SHAP analysis reveals clear differences in the conditions under which the main anthropogenic fire causes arise. Fires attributed to negligence align closely with the combination of enabling factors described in the literature: dry and windy weather, high evapotranspiration, and agricultural activity near flammable vegetation. These conditions suggest that only a minimal ignition source is needed for a fire to start and spread. Intentional fires, by contrast, tend to occur under cooler and wetter environmental conditions, indicating that they do not rely on favourable fire weather but are instead initiated independently of climatic triggers. This divergence in climatic context provides a meaningful temporal dimension for interpreting fire causes and aligns with the idea that human-caused ignitions respond differently to environmental constraints. A further distinction is reflected in burned area patterns: intentional fires are more often associated with larger fire events, whereas negligence-related fires tend to remain smaller. Hantson, Andela [42] suggests that human-caused ignitions can generate more extreme fire behaviour when they coincide—either unintentionally or deliberately—with periods of heightened fire weather conditions, allowing fires to scape initial suppression efforts. While the analysis by Hantson, Andela [42] treated all human ignitions as a single category, our results provide a more nuanced perspective by showing that intentional and negligent ignitions are associated with different climatic contexts and consequently lead to distinct fire outcomes.

https://doi.org/10.3390/fire9040138

Previous research has shown that lightning-caused fires tend to occur in remote areas with limited human presence, where suppression is less immediate, and their occurrence is strongly modulated by weather conditions [2,11,43,44]. Conditions such as moderate fuel moisture, intermediate productivity, and topographic features that influence storm development are known to shape the spatial patterns of these events. In line with this understanding, our results indicate that lightning-caused fires are associated with landscapes of low human density and limited infrastructure, occurring predominantly in areas influenced by WFI and under intermediate wind speed and evapotranspiration conditions. These fires also tend to be linked to larger burned areas, although the contribution of this variable to the model prediction is comparatively modest, suggesting that, while lightning ignitions may grow large, their spatial signature is shaped more by biophysical context than by fire size itself. Overall, the patterns captured by our model align with the broader characterisation of lightning-driven fire regimes: ignition emerging in sparsely populated, fuel-rich environments where atmospheric conditions are sufficient—but not necessarily extreme—to enable fire ignitions and spread.

While Bountzouklis, Fox [2] used ignition point coordinates as predictors to estimate overall variable importance, our analysis advances this approach by computing importance values at the observation level. This enables spatial mapping and reveals how the relevance of specific predictors varies across the study area when distinguishing among different fire types. Our results indicate that human-caused fires are concentrated primarily in the NW region, fires of natural origin occur more frequently in the E region, and fires of unknown cause are distributed across multiple regions, with a notable density in the NW—an area that also exhibits the highest overall fire incidence.

Local importance analyses further allow us to assess whether the model's predictive behaviour is spatially consistent with these patterns and provide insight into whether particular drivers exert stronger influence in some regions than in others. Despite its advantages, local interpretability analysis has seen limited application in wildfire cause modelling studies. Ochoa, Bar-Massada [11], who conducted a comparable analysis at the European scale, reported that climatic variables play a substantial role in shaping fire occurrence patterns in the Mediterranean. As shown in Figure 6d, the aggregated representation is strongly shaped by the predominance of intentional fire records in the dataset, which drives much of the overall signal. Consequently, finer spatial nuances characterising individual fire cause categories become partially obscured when observations are pooled at the regional scale, blurring the distinct environmental and socio-ecological mechanisms that operate within each fire cause. This illustrates the need for disaggregated or cause-specific analyses to fully resolve the spatial variability underpinning fire occurrence.

Interpreting causality patterns across datasets requires careful consideration of the methodological choices and data limitations that shape fire statistics. Differences observed between ESFire30 and EGIF become more evident once fire causes from EGIF are transferred to the ESFire30 dataset. Several methodological and data-related factors may explain these discrepancies. First, ESFire30 consistently reports a higher total number of fires than EGIF (Ochoa et al., 2026: under review), which inevitably results in a larger proportional of events classified as unknown. Second, EGIF shows a marked decline in data availability after 2016, potentially affecting the relative distribution of causes in recent years. Finally, the majority-based approach used to assign causes to the fire perimeters tends to emphasise dominant categories—most notably intentional and unknown—while reducing the representation of less frequent causes such as negligence or natural ignitions.

Overall, our findings demonstrate the importance of integrating detailed causal information, robust spatialisation procedures, and local interpretability tools to better understand the mechanisms driving different types of fire ignitions. By aligning model outputs

https://doi.org/10.3390/fire9040138

with well-established fire drivers and highlighting areas where uncertainties or data limitations persist, this work contributes to a more refined characterisation of fire causation across Peninsular Spain. The combination of global and local interpretability provides a more complete view of how climatic and human factors interact spatially to shape fire patterns. Together these methodological steps improve our ability to distinguish among ignition causes and provide a robust, spatially explicit basis for developing evidence-based fire risk management approaches.

5. Conclusions

This study demonstrates that integrating ignition information from EGIF with satellite-derived fire perimeters from ESFire30, combined with machine learning classification, provides an effective approach for exploring wildfire ignition causes in Peninsular Spain.

Although the approach involves several sources of uncertainty—especially for minority classes and years with limited data—the model produces spatial and environmental patterns that are consistent with the patterns shown in the EGIF. The use of SHAP values, both globally and locally, provides additional context for interpreting how different drivers relate to each ignition type, while also highlighting areas where data limitations remain important. Overall, the framework presented here is not intended to deliver causal explanations; rather, it provides a structure and spatially explicit way to integrate heterogeneous wildfire information and reduce uncertainties associated with fire cause attribution. By aligning the ESFire30 dataset with the EGIF reference distribution, improving the representation of previously unknown causes, and linking fire occurrences to distinct environmental and anthropogenic drivers, this approach reveals consistent regional patterns and contrasts between fire cause types—particularly between intentional and negligence-related fires, and between human and natural ignitions. The spatial and regional SHAP analyses further highlight how the relative importance of climatic, vegetation, and socio-territorial variables shifts across pyro-regions, while also showing the dominant influence of evapotranspiration and wind speed at the broader scale. In this way, the framework helps organise and interpret complex fire cause dynamics, contributing to a more coherent understanding of wildfire drivers in mainland Spain, and supporting a more targeted prevention and management planning.

In addition, this study builds upon and improves the ESFire30 dataset by integrating all relevant environmental and socio-economic attributes discussed in this paper, including land-use, road, grazing, and interface data. The dataset, available as shp (SHP), now links detailed burned area perimeters with these explanatory variables, enabling a more comprehensive analysis of fire patterns and causes. This enhanced dataset provides a valuable resource for developing robust models to identify predominant fire causes and supporting future research on fire dynamics and risk assessment in the region.

Author Contributions: Conceptualization, C.O., M.F., M.R. and E.C.; Methodology, C.O., M.F. and M.R.; Software, C.O.; Formal analysis, C.O.; Investigation, C.O., M.F. and E.C.; Data curation, C.O. and M.F.; Writing—original draft, C.O.; Writing—review & editing, C.O., M.F., M.R. and E.C.; Supervision, E.C.; Project administration, E.C.; Funding acquisition, E.C. All authors have read and agreed to the published version of the manuscript.

Funding: This paper is funded by the H2020 Project FirEUrisk (Grant Agreement Number: 101003890). The FirEUrisk project contributions reported in this paper were carried out by Rodrigues, M., Ochoa, C. and Chuvieco, E. Franquesa, M. was supported by the Grant JDC2022-048710-I, funded by MCIN/AEI/10.13039/501100011033, and by the European Union NextGenerationEU/PRTR.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

https://doi.org/10.3390/fire9040138

Data Availability Statement: The complete database used in this study has been made publicly available and is accessible through the following DOI: https://doi.org/10.5281/zenodo.18449006.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Costafreda-Aumedes, S.; Comas, C.; Vega-Garcia, C. Human-caused fire occurrence modelling in perspective: A review. *Int. J. Wildland Fire* **2017**, *26*, 983. [CrossRef]
2. Bountzouklis, C.; Fox, D.M.; Bernardino, E.D. Predicting wildfire ignition causes in Southern France using eXplainable Artificial Intelligence (XAI) methods. *Environ. Res. Lett.* **2023**, *18*, 044038. [CrossRef]
3. Jones, M.W.; Abatzoglou, J.T.; Veraverbeke, S.; Andela, N.; Lasslop, G.; Forkel, M.; Smith, A.J.; Burton, C.; Betts, R.A.; van der Werf, G.R. Global and regional trends and drivers of fire under climate change. *Rev. Geophys.* **2022**, *60*, e2020RG000726. [CrossRef]
4. Martinez, J.; Vega-Garcia, C.; Chuvieco, E. Human-caused wildfire risk rating for prevention planning in Spain. *J. Environ. Manag.* **2009**, *90*, 1241–1252. [CrossRef]
5. Rodrigues, M.; Jiménez-Ruano, A.; Peña-Angulo, D.; de la Riva, J. A comprehensive spatial-temporal analysis of driving factors of human-caused wildfires in Spain using Geographically Weighted Logistic Regression. *J. Environ. Manag.* **2018**, *225*, 177–192. [CrossRef]
6. Pineda, N.; Rigo, T. The rainfall factor in lightning-ignited wildfires in Catalonia. *Agric. For. Meteorol.* **2017**, *239*, 249–263. [CrossRef]
7. Chicas, S.D.; Nielsen, J.Ø. Who are the actors and what are the factors that are used in models to map forest fire susceptibility? A systematic review. *Nat. Hazards* **2022**, *114*, 2417–2434. [CrossRef]
8. Rodrigues, M.; la Riva, J.D. Assessing the effect on fire risk modeling of the uncertainty in the location and cause of forest fires. In *Advances in Forest Fire Research*; International Association of Wildland Fire (IAWF): Missoula, MT, USA, 2014; pp. 1061–1072.
9. Lundberg, S.M.; Lee, S.-I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 4768–4777.
10. Jain, P.; Coogan, S.C.P.; Subramanian, S.G.; Crowley, M.; Taylor, S.; Flannigan, M.D. A review of machine learning applications in wildfire science and management. *Environ. Rev.* **2020**, *28*, 478–505. [CrossRef]
11. Ochoa, C.; Bar-Massada, A.; Chuvieco, E. A European-scale analysis reveals the complex roles of anthropogenic and climatic factors in driving the initiation of large wildfires. *Sci. Total Environ.* **2024**, *917*, 170443. [CrossRef] [PubMed]
12. Ochoa, C.; Khairoun, A.; Coello, J.A.R.; Chuvieco, E. ESFire30. 2026. Available online: https://zenodo.org/records/18451006 (accessed on 1 January 2026).
13. Quan, Y.; Zhong, X.; Feng, W.; Chan, J.C.-W.; Li, Q.; Xing, M. SMOTE-Based Weighted Deep Rotation Forest for the Imbalanced Hyperspectral Data Classification. *Remote Sens.* **2021**, *13*, 464. [CrossRef]
14. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
15. Rodrigues, M.; Jiménez, A.; de la Riva, J. Analysis of recent spatial–temporal evolution of human driving factors of wildfires in Spain. *Nat. Hazards* **2016**, *84*, 2049–2070. [CrossRef]
16. Rodrigues, M.; de la Riva, J. An insight into machine-learning algorithms to model human-caused wildfire occurrence. *Environ. Model. Softw.* **2014**, *57*, 192–201. [CrossRef]
17. Chuvieco, E.; Aguado, I.; Jurdao, S.; Pettinari, M.L.; Yebra, M.; Salas, J.; Hantson, S.; de la Riva, J.; Ibarra, P.; Rodrigues, M. Integrating geospatial information into fire risk assessment. *Int. J. Wildland Fire* **2014**, *23*, 606–619. [CrossRef]
18. Nunes, A.; Lourenço, L.; Meira, A.C. Exploring spatial patterns and drivers of forest fires in Portugal (1980–2014). *Sci. Total Environ.* **2016**, *573*, 1190–1202. [CrossRef]
19. Nunes, A.N. Regional variability and driving forces behind forest fires in Portugal an overview of the last three decades (1980–2009). *Appl. Geogr.* **2012**, *34*, 576–586. [CrossRef]
20. Padilla, M.; Vega-García, C. On the comparative importance of fire danger rating indices and their integration with spatial and temporal variables for predicting daily human-caused fire occurrences in Spain. *Int. J. Wildland Fire* **2011**, *20*, 46–58. [CrossRef]
21. Leone, V.; Lovreglio, R.; Martín, M.P.; Martínez, J.; Vilar, L. Human factors of fire occurrence in the Mediterranean. In *Earth Observation of Wildland Fires in Mediterranean Ecosystems*; Springer: Berlin/Heidelberg, Germany, 2009; pp. 149–170.
22. Leone, V.; Koutsias, N.; Martínez, J.; Vega-García, C.; Allgöwer, B.; Lovreglio, R. The human factor in fire danger assessment. In *Wildland Fire Danger Estimation and Mapping: The Role of Remote Sensing Data*; World Scientific: Singapore, 2003; pp. 143–196.
23. Escolano, C.L.; Campos, Á.P.; Vidal, R.P.; Logroño, M.P.A. Valoración y representación cartográfica de la accesibilidad viaria en la España peninsular: 1960–2014. *GeoFocus* **2016**, *18*, 169–189.
24. Bar-Massada, A.; Stewart, S.I.; Hammer, R.B.; Mockrin, M.H.; Radeloff, V.C. Using structure locations as a basis for mapping the wildland urban interface. *J. Environ. Manag.* **2013**, *128*, 540–547. [CrossRef] [PubMed]

https://doi.org/10.3390/fire9040138

25. Bar-Massada, A. A comparative analysis of two major approaches for mapping the wildland-urban interface: A case study in California. *Land* **2021**, *10*, 679. [CrossRef]
26. Pesaresi, M.; Schiavina, M.; Politis, P.; Freire, S.; Krasnodębska, K.; Uhl, J.H.; Carioli, A.; Corbane, C.; Dijkstra, L.; Florio, P. Advances on the Global Human Settlement Layer by joint assessment of Earth Observation and population survey data. *Int. J. Digit. Earth* **2024**, *17*, 2390454. [CrossRef]
27. WoldPop. *Population Density—WoldPop Project Categories*; University of Southampton, WorldPop—School of Geography and Environmental Science: Southampton, UK, 2025.
28. Ministerio de Agricultura, Pesca y Alimentación. *Demografía en la Población Rural en España in Análisis y Prospectiva 2020*; Ministerio de Agricultura, Pesca y Alimentación: Madrid, Spain, 2020.
29. European Environment Agency. *Net Greenhouse Gas Emissions of the LULUCF Sector*; European Environment Agency: Copenhagen, Denmark, 2025.
30. Serrano-Zulueta, R.; Gómez-Sal, A.; Pauné, F.; Velado-Alonso, E.; Garzón, J.; del Prado, A.; Herrera, P.M.; Majadas, J.; Pasetti, F.; Prada-Llorente, E.; et al. A Classification of Pastoralism in Spain: Understanding the Past to Address Present Challenges. *Nomadic Peoples* **2024**, *28*, 242–274. [CrossRef]
31. Acebes, P.; Iglesias-González, Z.; Muñoz-Galvez, F.J. Do Traditional Livestock Systems Fit into Contemporary Landscapes? Integrating Social Perceptions and Values on Landscape Change. *Agriculture* **2021**, *11*, 1107. [CrossRef]
32. Vicente-Serrano, S.M.; Tomas-Burguera, M.; Beguería, S.; Reig, F.; Latorre, B.; Peña-Gallardo, M.; Luna, M.Y.; Morata, A.; González-Hidalgo, J.C. A high resolution dataset of drought indices for Spain. *Data* **2017**, *2*, 22. [CrossRef]
33. Vicente-Serrano, S.M.; Beguería, S.; López-Moreno, J.I. A Multiscalar Drought Index Sensitive to Global Warming: The Standardized Precipitation Evapotranspiration Index. *J. Clim.* **2010**, *23*, 1696–1718. [CrossRef]
34. Franquesa, M.; Reig-Gracia, F.; Serrano, S.M.V. *Vegetation Indices for the Iberian Peninsula and Balearic Islands (VIIB) Database*; Universidad de Zaragoza: Zaragoza, Spain, 2024.
35. Stevens, F.R.; Gaughan, A.E.; Linard, C.; Tatem, A.J. Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data. *PLoS ONE* **2015**, *10*, e0107042. [CrossRef]
36. IGN. *Redes de Transporte del CNIG*; Instituto Geográfico Nacional (IGN): Madrid, Spain, 2024.
37. Centro Nacional de Información Geográfica (CNIG). *Sistema de Información sobre la Ocupación del Suelo en España (SIOSE): Series 2005–2020*; Ministerio de Transportes y Movilidad Sostenible, Gobierno de España: Madrid, Spain, 2020.
38. European Environment Agency (EEA). *CORINE Land Cover (CLC) Dataset Series: 1990, 2000, 2006, 2012, 2018*; European Environment Agency: Copenhagen, Denmark, 2018.
39. Gilbert, M.; Nicolas, G.; Cinardi, G.; Van Boeckel, T.P.; Vanwambeke, S.O.; Wint, G.R.W.; Robinson, T.P. Global distribution data for cattle, buffaloes, horses, sheep, goats, pigs, chickens and ducks in 2010. *Sci. Data* **2018**, *5*, 180227. [CrossRef]
40. Franquesa, M.; Reig, F.; Arretxea, M.; Adell-Michavila, M.; Halifa-Marín, A.; Vilas, D.; Beguería, S.; Vicente-Serrano, S.M. Near-real-time vegetation monitoring and historical database (1981–present) for the Iberian Peninsula and the Balearic Islands. *Earth Syst. Sci. Data* **2025**, *17*, 5885–5902. [CrossRef]
41. Trigo, R.M.; Sousa, P.M.; Pereira, M.G.; Rasilla, D.; Gouveia, C.M. Modelling wildfire activity in Iberia with different atmospheric circulation weather types. *Int. J. Clim.* **2016**, *36*, 2761–2778. [CrossRef]
42. Hantson, S.; Andela, N.; Goulden, M.L.; Randerson, J.T. Human-ignited fires result in more extreme fire behavior and ecosystem impacts. *Nat. Commun.* **2022**, *13*, 2717. [CrossRef] [PubMed]
43. Rodrigues, M.; Jiménez-Ruano, A.; Gelabert, P.J.; de Dios, V.R.; Torres, L.; Ribalaygua, J.; Vega-García, C. Modelling the daily probability of lightning-caused ignition in the Iberian Peninsula. *Int. J. Wildland Fire* **2023**, *32*, 351–362. [CrossRef]
44. Johnson, E.; Miyanishi, K.; Bridge, S. Wildfire regime in the boreal forest and the idea of suppression and fuel buildup. *Conserv. Biol.* **2001**, *15*, 1554–1557. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

https://doi.org/10.3390/fire9040138

remote sensing

MDPI

Article

Interval-Based Tropical Cyclone Intensity Forecasting with Spatiotemporal Transformers

Tao Guo [1,2,†], Hua Zhang [3,4,†], Tao Song [1,2] and Shiqiu Peng [3,4,5,*]

[1] Qingdao Institute of Software, College of Computer Science and Technology, China University of Petroleum (East China), Qingdao 266580, China; s23070039@s.upc.edu.cn (T.G.)
[2] Shandong Key Laboratory of Intelligent Oil & Gas Industrial Software, China University of Petroleum (East China), Qingdao 266580, China
[3] State Key Laboratory of Tropical Oceanography, South China Sea Institute of Oceanology, Chinese Academy of Sciences, Guangzhou 510301, China
[4] Guangdong Key Laboratory of Ocean Remote Sensing and Big Data, South China Sea Institute of Oceanology, Chinese Academy of Sciences, Guangzhou 510301, China
[5] Laboratory for Regional Oceanography and Numerical Modeling, Qingdao Marine Science and Technology Center, Qingdao 266237, China
* Correspondence: speng@scsio.ac.cn
† These authors contributed equally to this work.

Highlights

What are the main findings?

- A unified probabilistic transformer framework (TC-QFormer) bridges spatiotemporal satellite representation learning and tropical cyclone intensity forecasting via a physically informed scalar–image fusion mechanism.
- The framework enables interval-based probabilistic forecasting while preserving deterministic prediction accuracy, thereby improving the reliability of tropical cyclone intensity prediction.

What are the implications of the main findings?

- The framework supports uncertainty-aware tropical cyclone intensity forecasting, contributing to operational risk assessment and early-warning decision-making.
- Integrating physically informed conditioning within probabilistic transformer models offers a scalable and systematic pathway for future tropical cyclone intensity forecasting systems.

Abstract

Accurate tropical cyclone (TC) intensity forecasting remains challenging due to the strong nonlinearity of intensity evolution and the rapid structural changes associated with storm development. In this work, we propose TC-QFormer, an interval-based probabilistic framework for 24 h TC intensity forecasting that combines transformer-based spatiotemporal modeling with scalar conditioning. Specifically, we adapt the PredFormer video prediction model for multi-horizon scalar regression and introduce a lightweight Scalar–Image Fusion Block to incorporate historical intensity information into the visual representations. A two-stage training strategy is adopted, in which the model is first pretrained for deterministic median prediction and subsequently fine-tuned to directly predict multiple conditional quantiles using the pinball loss. Experiments are conducted on the TCIR dataset using geostationary infrared and water vapor satellite imagery together with aligned historical intensity records. The proposed method is evaluated against representative recurrent and non-recurrent baselines, including ConvLSTM, PredRNN, and SimVP. Results indicate that

the proposed framework achieves improved deterministic accuracy and produces well-calibrated 80% prediction intervals, particularly at longer forecast lead times and during rapidly evolving intensity regimes. These findings suggest that combining transformer-based spatiotemporal modeling with scalar–image conditioning provides an effective and interpretable approach for probabilistic TC intensity forecasting.

Keywords: tropical cyclone intensity forecasting; probabilistic forecasting; quantile regression; spatiotemporal transformer; scalar–image fusion

1. Introduction

Tropical cyclones (TCs) are among the most destructive natural hazards worldwide, causing severe economic losses and societal impacts [1–3]. While substantial progress has been made in predicting TC tracks, forecasting the intensity evolution of TCs remains considerably more challenging, particularly on short to medium time scales [4,5]. This difficulty arises from the inherently nonlinear interactions among atmospheric circulation, ocean heat content, air–sea fluxes, and storm inner-core processes [6–8]. Operational TC intensity prediction has traditionally relied on statistical–dynamical guidance models, such as the Statistical Hurricane Intensity Prediction Scheme (SHIPS), which combine climatological relationships with environmental predictors derived from numerical weather prediction outputs [9,10]. In parallel, satellite-based intensity estimation techniques, including the Dvorak method [11] and its automated variants [12], have played a crucial role in providing objective estimates of TC intensity from geostationary observations. Despite continuous improvements, these approaches remain limited in their ability to anticipate rapid intensity changes and to provide reliable uncertainty information, particularly during rapid intensification events and at longer lead times [13,14]. This limitation motivates the exploration of data-driven methods that can directly exploit the evolving storm structure observed in satellite imagery.

In recent years, deep learning has emerged as a powerful tool for TC intensity estimation and forecasting that works by leveraging large volumes of satellite imagery and historical best-track data [15,16]. Early studies primarily employed convolutional neural networks (CNNs) to estimate intensity from single satellite frames, demonstrating clear improvements over traditional satellite analysis techniques [15,16]. To better capture temporal evolution, recurrent architectures such as convolutional long short-term memory networks (ConvLSTM) [17] and PredRNN-style models [18] were subsequently introduced, adapting video prediction frameworks to meteorological applications. These models highlight an important insight: the temporal evolution of storm structure, rather than isolated snapshots, plays a central role in intensity change [17,18]. More recently, simplified non-recurrent spatiotemporal models such as SimVP [19] have achieved competitive performance with reduced computational complexity, further demonstrating the effectiveness of spatiotemporal modeling for TC intensity forecasting.

Attention-based transformer architectures have become increasingly influential for long-term time series forecasting due to their ability to model long-range dependencies [20,21]. Variants such as Informer [22], Autoformer [23], and TimesNet [24] have been widely explored for multivariate time series prediction. In the spatiotemporal domain, transformer-based models such as PredFormer [25] extend this capability by integrating temporal and spatial attention within a unified framework, offering a flexible backbone for video prediction and meteorological forecasting. Beyond image-based approaches, transformers have also been applied to TC track and intensity forecasting using historical storm

https://doi.org/10.3390/rs18071069

attributes and environmental predictors [26]. These studies suggest that transformer-based spatiotemporal models are well suited to capture the global and multiscale structure–intensity relationships inherent in TCs.

Despite advances in deterministic forecasting, there is growing recognition that an explicit representation of uncertainty is essential for operational TC intensity prediction and risk assessment [27]. From a statistical perspective, probabilistic forecasting provides a principled framework for characterizing uncertainty and supporting decision-making under risk [28]. In meteorology, uncertainty has traditionally been addressed through ensemble-based approaches and Bayesian formulations [29,30].

An alternative and computationally efficient approach to probabilistic forecasting is quantile regression, which directly estimates conditional quantiles of the target variable without assuming a full probability density function [31,32]. Quantile-based methods have been widely adopted in energy and wind power forecasting, where deep neural networks optimized with the pinball loss have demonstrated strong performance [33,34]. Recent studies have further documented the growing adoption of quantile regression across data-driven probabilistic forecasting applications [35]. Compared with ensemble or generative models, quantile-based methods offer a practical compromise by producing interval-based forecasts that are both interpretable and operationally meaningful, making them particularly attractive for TC intensity forecasting.

In the context of TC intensity forecasting, probabilistic approaches have attracted increasing attention in recent years. Ensemble-based intensity guidance remains widely used in operational practice [14,36], while more recent studies have explored generative models, such as diffusion-based frameworks, to estimate full predictive distributions of TC intensity [37]. Although these methods provide a comprehensive probabilistic description, they often involve increased model complexity and substantial computational resources. In contrast, interval-based probabilistic forecasts derived from quantile regression remain relatively underexplored, particularly in combination with spatiotemporal modeling of satellite imagery [31,32].

This study proposes an interval-based probabilistic framework for short-term TC intensity forecasting that directly leverages the spatiotemporal evolution of satellite-observed storm structures. We employ PredFormer as a transformer-based spatiotemporal backbone and extend it to multi-horizon scalar regression through a quantile-based formulation [32]. To further incorporate physically meaningful prior information, we introduce a lightweight Scalar–Image Fusion Block (SIFB) that conditions latent visual representations on historical intensity and auxiliary storm attributes. Rather than estimating a full predictive distribution, the proposed approach focuses on delivering interpretable prediction intervals alongside the median forecast. Using the TC for Image-to-Intensity Regression (TCIR) dataset [15], we systematically evaluate the proposed framework against representative recurrent and non-recurrent baselines, including ConvLSTM, PredRNN, and SimVP, and analyze its performance across different forecast lead times, including rapid intensification events [13].

2. Data

2.1. TCIR Dataset

This study uses the Tropical Cyclone for Image-to-Intensity Regression (TCIR) dataset [38], which provides temporally continuous geostationary satellite imagery collocated with best-track intensity records for 1285 TCs during the period 2003–2016 across all major ocean basins. The target variable is the maximum sustained wind speed ($V_{\max}$, kt), following standard best-track conventions [39]. In addition to satellite imagery, TCIR provides several storm-level scalar variables (Table 1), including maximum sustained wind

https://doi.org/10.3390/rs18071069

speed (V_{max}), minimum sea-level pressure (MSLP), mean radius of 35 kt winds across four quadrants (R35_4qAVG), and storm center location (Lat and Lon). In this study, Vmax serves as the prediction target, while different subsets of scalar variables are optionally used as auxiliary inputs for scalar–image fusion experiments.

Table 1. Scalar variables provided in the TCIR dataset and their usage in this study.

Variable	Description	Unit	Usage
Vmax	Maximum sustained wind speed	kt	Target/Input
MSLP	Minimum sea-level pressure	hPa	Optional input
R35_4qAVG	Mean radius of 35 kt winds (4 quadrants)	nmi	Optional input
Lat	TC center latitude	degree	Optional input
Lon	TC center longitude	degree	Optional input

2.2. Satellite Channels and Spatial Configuration

The TCIR dataset includes multiple satellite channels, including geostationary infrared (IR), water vapor (WV), visible (VIS), and passive microwave (PMW). In this study, only IR and WV imagery are used as primary inputs. These channels provide continuous day–night coverage at regular 3 h temporal resolution, making them well suited for consistent spatiotemporal modeling of TC structure. In contrast, VIS is excluded due to the absence of nighttime samples, and PMW is not used because of its irregular temporal sampling and coarser effective resolution. Raw images are cropped to a 201×201 region centered on the storm and resized to 128×128 using bilinear interpolation.

2.3. Temporal Configuration

Satellite observations in TCIR are available at 3 h intervals. As shown in Figure 1, for each prediction instance, the model ingests a 12 h historical window consisting of four frames at $T-9$, $T-6$, $T-3$, and T hours, with IR and WV concatenated as a two-channel input. The forecasting target is an 8-step, 24 h prediction from $T+3$ to $T+24$ h at 3 h intervals.

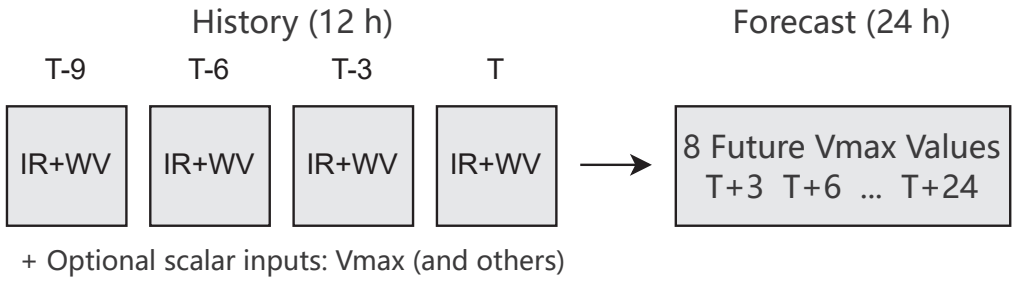


Figure 1. Temporal forecasting configuration in this study. Four historical IR/WV frames (12 h history, $T-9$ to T) are used to predict eight future V_{max} values (24 h forecast, $T+3$ to $T+24$) at 3 h intervals, with optional scalar inputs.

2.4. Preprocessing

Satellite images are first screened for missing or invalid values. To avoid interpolation artifacts, any NaN or infinite pixel values within an input window are replaced with zeros prior to resizing. These values serve only as placeholders for invalid pixels and do not represent meaningful physical measurements in the input data. The cropped 201×201 images are then resized to 128×128 using bilinear interpolation. This resolution provides a practical balance between computational efficiency and the preservation of the large-scale structural features of tropical cyclones, such as the organization of the inner core and surrounding cloud bands. In addition, the 128×128 resolution aligns well

https://doi.org/10.3390/rs18071069

with common deep learning model architectures and patch-based processing schemes, enabling efficient tensor operations and stable model training while retaining sufficient spatial information for short-term intensity forecasting.

Image inputs are normalized using a per-sample z-score scheme applied independently to each time step and channel over spatial pixels. A minimum standard deviation threshold is imposed for numerical stability, and normalized values are clipped to $[-6, 6]$. Scalar variables are standardized using statistics computed from the training set only, with the same parameters applied to validation and test data to prevent information leakage. Samples with invalid target records are excluded during dataset construction.

2.5. Train–Validation–Test Split and Dataset Statistics

A storm-level chronological split is adopted to ensure realistic evaluation and prevent information leakage across datasets. Each tropical cyclone is identified by a unique storm key, and all samples from the same storm are assigned to a single subset based on the first year of its records. Specifically, storms first recorded during 2003–2013 are used for training, storms from 2014 are used for validation, and storms from 2015 to 2016 are used for testing. Prediction instances are constructed using a sliding window along each storm track, and input windows with a missing-value ratio exceeding 20% are excluded during index construction.

As summarized in Table 2, the final dataset contains 1005 training storms (43,779 samples), 81 validation storms (3843 samples), and 179 test storms (8426 samples), corresponding to 350,232, 30,744, and 67,408 target intensity values, respectively. Rapid intensification (RI) events are defined as cases where the maximum sustained wind speed increases by at least 30 kt within 24 h. The RI ratios are 6.30%, 7.42%, and 8.18% for the training, validation, and test sets, respectively.

The dataset includes tropical cyclones from all major ocean basins. The basin-wise distribution of storms and samples is summarized in Table 3. The largest proportions come from the western Pacific and the Southern Hemisphere, followed by the eastern Pacific and the North Atlantic, while the Indian Ocean and central Pacific comprise smaller portions of the dataset.

Table 2. Dataset split statistics for TC intensity forecasting. RI denotes rapid intensification events defined as cases with a 24 h increase in maximum sustained wind speed of at least 30 kt.

Split	#Storms	#Samples	Year Range	#Intensity Targets	#RI Samples	RI Ratio
Train	1005	43,779	2003–2013	350,232	2756	6.30%
Val	81	3843	2014	30,744	285	7.42%
Test	179	8426	2015–2016	67,408	689	8.18%

Table 3. Basin-wise distribution of storms and samples in the dataset.

Basin	#Storms	#Samples	Sample Ratio
WP	371	15,877	28.33%
EP	242	10,890	19.43%
CP	19	1270	2.27%
NA	228	10,889	19.43%
IO	75	2376	4.24%
SH	330	14,746	26.31%

https://doi.org/10.3390/rs18071069

3. Methodology

3.1. Overall Framework

We propose an interval-based probabilistic framework for TC intensity forecasting, termed **TC-QFormer**, which integrates spatiotemporal transformer modeling with scalar-conditioned latent fusion. Given historical sequences of IR–WV satellite imagery and aligned intensity records, the framework learns compact spatiotemporal representations of TC structural evolution and directly outputs non-autoregressive, multi-horizon quantile forecasts of future intensity.

As illustrated in Figure 2, the proposed framework consists of four main components: (i) a spatiotemporal transformer backbone adapted from PredFormer [25], (ii) a SIFB for conditioning latent representations on historical scalar inputs, (iii) a joint non-autoregressive quantile regression head, and (iv) a pinball-loss-based training strategy for probabilistic intensity forecasting.

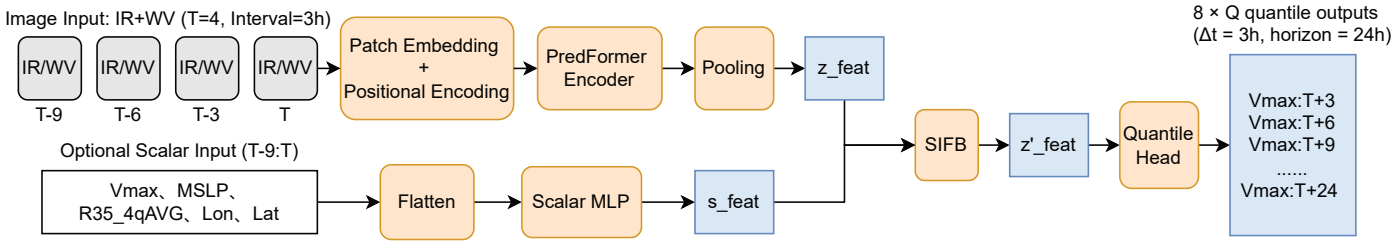


Figure 2. Overall framework of TC-QFormer model. Historical IR/WV sequences ($T = 4$, 3 h interval) are encoded by a PredFormer backbone to produce a pooled latent feature $\mathbf{z}$. Optional scalar histories are embedded into $\mathbf{s}$ and fused with $\mathbf{z}$ via a Scalar–Image Fusion Block (SIFB) to obtain $\mathbf{z}'$. A joint non-autoregressive quantile head outputs multi-horizon forecasts ($8 \times |\mathcal{Q}|$; 24 h horizon).

Let

$$\mathbf{X} = \{\mathbf{x}_{t-T+1}, \ldots, \mathbf{x}_t\}, \quad \mathbf{x}_i \in \mathbb{R}^{H \times W \times C} \tag{1}$$

denote the sequence of T historical IR–WV satellite frames (with $T = 4$ in this study), and let

$$\mathbf{s} = \{s_{t-T+1}, \ldots, s_t\} \tag{2}$$

denote the corresponding scalar intensity sequence (e.g., maximum sustained wind speed in kt).

The objective is to predict T_{out} conditional intensity quantiles at each future lead time,

$$\left\{\hat{y}_{t+\tau}^{(q)}\right\}_{\tau=1}^{T_{\text{out}}}, \quad q \in \mathcal{Q}, \tag{3}$$

where $\mathcal{Q} = \{0.1, 0.25, 0.5, 0.75, 0.9\}$ and $T_{\text{out}} = 8$ in our experiments.

3.2. Encoder and Scalar Fusion Architecture

Figure 3 presents the detailed architecture of the proposed encoder-only backbone and the scalar–image fusion module.

3.2.1. PredFormer Encoder

To model the spatiotemporal evolution of TCs, we adopt PredFormer as the visual backbone and retain only its encoder component (Figure 3a).

https://doi.org/10.3390/rs18071069

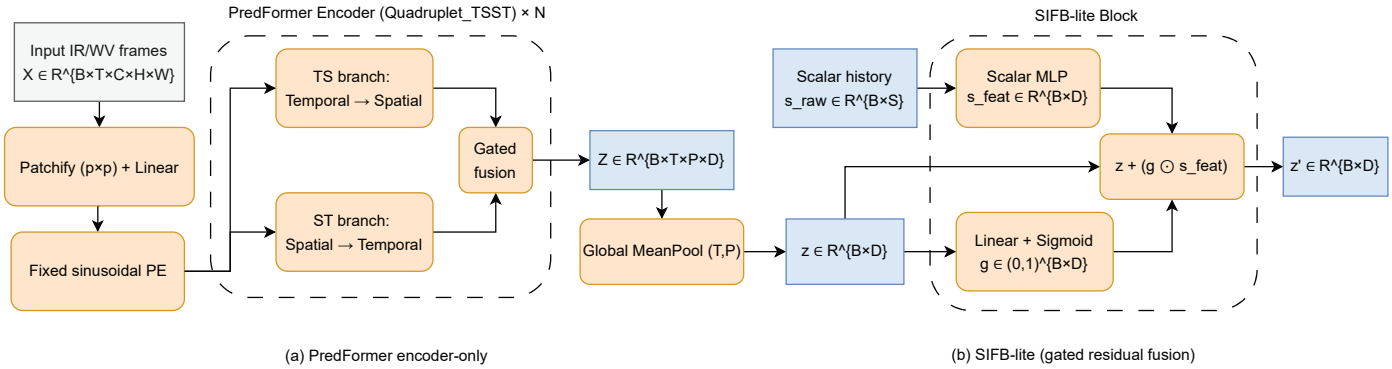


Figure 3. Detailed architecture of the proposed encoder and scalar fusion module. (**a**) Encoder-only adaptation of PredFormer with Quadruplet_TSST blocks. Input IR/WV sequences are patch-embedded with fixed sinusoidal positional encoding, processed by N stacked PredFormer layers, and globally pooled over temporal and spatial tokens to obtain a compact latent representation $\mathbf{z} \in \mathbb{R}^{B\times D}$. (**b**) Scalar–Image Fusion Block (SIFB-lite). Historical scalar inputs are embedded into $\mathbf{s}_{\text{feat}}$ via an MLP. A gating vector $\mathbf{g} = \sigma(\mathbf{W}\mathbf{z})$ modulates scalar injection, and a residual formulation $\mathbf{z}' = \mathbf{z} + \mathbf{g} \odot \mathbf{s}_{\text{feat}}$ produces the fused representation for downstream quantile prediction.

Given an input sequence $\mathbf{X} \in \mathbb{R}^{B\times T\times C\times H\times W}$ where $C = 2$ corresponds to IR and WV channels, each frame is partitioned into non-overlapping patches of size $p \times p$ ($p = 8$). After flattening and linear projection, patch tokens are obtained as

$$\mathbf{Z}_0 \in \mathbb{R}^{B\times T\times P\times D}, \tag{4}$$

where $P = (H/p)(W/p)$ denotes the number of patches per frame. A fixed sinusoidal spatiotemporal positional embedding is added to provide explicit indexing over temporal and spatial dimensions.

The token sequence is then processed by N_{depth} stacked PredFormer layers:

$$\mathbf{Z}_{\ell+1} = f_\ell(\mathbf{Z}_\ell), \quad \ell = 0, \ldots, N_{\text{depth}} - 1. \tag{5}$$

Each layer adopts the Quadruplet_TSST interleaving strategy, which alternates temporal and spatial self-attention in both temporal-to-spatial (TS) and spatial-to-temporal (ST) branches, enabling efficient global spatiotemporal interaction without recurrent decoding.

Unlike the original video prediction setting, we discard the reconstruction decoder and employ an encoder-only formulation. The final token tensor $\mathbf{Z} \in \mathbb{R}^{B\times T\times P\times D}$ is globally averaged over temporal and spatial dimensions:

$$\mathbf{z} = \text{GlobalAvgPool}_{(T,P)}(\mathbf{Z}) \in \mathbb{R}^{B\times D}. \tag{6}$$

The pooled representation $\mathbf{z}$ serves as a compact latent descriptor for downstream intensity prediction.

3.2.2. Scalar–Image Fusion Block

Historical intensity and auxiliary scalar variables provide complementary physical context that may not be fully captured by imagery alone. We therefore introduce a lightweight SIFB that conditions the pooled latent representation on scalar inputs (Figure 3b). In experiments without auxiliary scalar inputs, SIFB reduces to identity mapping.

Let $\mathbf{s} \in \mathbb{R}^{B\times S}$ denote the flattened historical scalar sequence, where $S = T \times K_{\text{scalar}}$. It is embedded into the same latent dimension via a two-layer MLP:

$$\mathbf{s}_{\text{feat}} = \text{MLP}(\mathbf{s}) \in \mathbb{R}^{B\times D}. \tag{7}$$

https://doi.org/10.3390/rs18071069

A gating vector is computed from the visual representation to adaptively control scalar injection strength:

$$\mathbf{g} = \sigma(\mathbf{W}_g \mathbf{z} + \mathbf{b}_g), \tag{8}$$

and scalar information is injected through a residual gated formulation:

$$\mathbf{z}' = \mathbf{z} + \mathbf{g} \odot \mathbf{s}_{\text{feat}}. \tag{9}$$

This residual structure preserves the original visual representation while enabling adaptive scalar modulation. The fused latent feature $\mathbf{z}'$ is then fed to the quantile regression head for multi-horizon intensity forecasting.

3.3. Quantile-Based Probabilistic Intensity Prediction

Given the fused latent representation $\mathbf{z}' \in \mathbb{R}^{B \times D}$, we employ a lightweight multilayer perceptron to generate multi-horizon quantile forecasts in a joint non-autoregressive manner.

Let T_{out} denote the number of forecast lead times and $\mathcal{Q}$ the set of quantile levels. In our experiments, $T_{\text{out}} = 8$. The quantile head maps $\mathbf{z}'$ to

$$\hat{\mathbf{y}} = h(\mathbf{z}') \in \mathbb{R}^{B \times T_{\text{out}} \times |\mathcal{Q}|}, \tag{10}$$

where

$$h(\mathbf{z}') = \text{Reshape}_{T_{\text{out}},|\mathcal{Q}|}\big(\mathbf{W}_2\, \phi(\mathbf{W}_1 \mathbf{z}' + \mathbf{b}_1) + \mathbf{b}_2\big), \tag{11}$$

and $\phi(\cdot)$ denotes the GELU activation. All lead times and quantiles are predicted simultaneously from the same latent feature, avoiding iterative autoregressive rollouts and the associated error propagation across forecast steps.

In Stage 2, we use $\mathcal{Q} = \{0.1, 0.25, 0.5, 0.75, 0.9\}$, allowing direct construction of central prediction intervals such as 10–90%. In Stage 1, only the median ($q = 0.5$) is predicted, which is equivalent to ℓ_1 regression up to a constant factor. This two-stage strategy stabilizes deterministic learning before probabilistic fine-tuning.

4. Experimental Setup

4.1. Task Definition and Data Configuration

Unless otherwise specified, only IR and WV satellite imagery are used as visual input. At each time step, the two channels are concatenated along the channel dimension and resized to a fixed spatial resolution before being fed into the model. This IR–WV configuration provides continuous, all-weather, day–night coverage and serves as the default visual input throughout the experiments. In addition to satellite imagery, we consider the use of historical scalar storm attributes to condition intensity forecasts. Three scalar configurations are evaluated in the main experiments: (i) image-only input, where the model relies solely on IR and WV imagery; (ii) intensity-only input, where the historical V_{max} sequence is incorporated; and (iii) multi-scalar input, where five scalar storm attributes (V_{max}, MSLP, R35_4qAVG, Lat, and Lon) are used. For models equipped with scalar fusion, these scalar sequences are injected into the visual backbone via the proposed SIFB.

To explicitly assess the predictive value of satellite imagery, we additionally construct scalar-only baselines that operate exclusively on historical V_{max}. These scalar baselines include persistence-based extrapolation and a simple regression model, and are treated as a separate experimental group rather than as visual forecasting baselines. Their role is to provide a reference for quantifying the performance gap between purely scalar-based extrapolation and image-driven spatiotemporal forecasting.

https://doi.org/10.3390/rs18071069

4.2. Baseline Models

To comprehensively evaluate the proposed framework, we compare with representative spatiotemporal forecasting baselines that cover three mainstream modeling paradigms in video prediction and meteorological forecasting: recurrent neural networks (RNNs), CNNs, and transformer-based architectures:

- **ConvLSTM** [17]: A convolutional recurrent model that integrates convolutional operations into long short-term memory units to jointly capture spatial patterns and temporal dependencies. ConvLSTM represents the class of traditional RNN-based models that have been widely adopted in meteorological nowcasting and spatiotemporal sequence prediction.
- **PredRNN** [18]: An advanced recurrent architecture that introduces spatiotemporal memory units to better model long-range temporal dependencies. PredRNN serves as a strong autoregressive baseline within the RNN family for spatiotemporal predictive learning.
- **SimVP** [19]: A recurrent-free spatiotemporal prediction framework based on convolutional encoder–decoder structures. By removing recurrent rollouts, SimVP achieves competitive performance with improved computational efficiency, representing modern CNN-based video prediction models.
- **PredFormer** [25]: A pure vision transformer (ViT)-based spatiotemporal predictive model that relies solely on patch embeddings and attention mechanisms, without autoregressive recurrence or traditional convolution. PredFormer represents transformer-based forecasting paradigms that are well suited for modeling long-range spatiotemporal dependencies and scaling to more flexible input modalities.

All baseline models are adapted to the same input–output configuration as the proposed method. Specifically, they take satellite image sequences from the past 12 h (IR and WV channels) as input and predict TC intensity at eight future lead times over the subsequent 24 h. For fairness of comparison, ConvLSTM and PredRNN are evaluated using image-only inputs, while scalar conditioning is only enabled for PredFormer and SimVP through the proposed fusion mechanism.

Backbone selection rationale. Both SimVP and PredFormer achieve strong deterministic performance under identical settings. In this work, we adopt PredFormer as the primary backbone due to its stable multi-horizon predictions and compatibility with non-autoregressive quantile regression, which facilitates compact and well-calibrated uncertainty intervals.

4.3. Training Objective

The proposed framework produces interval-based probabilistic forecasts via quantile regression. We adopt the pinball loss as the training objective and apply a unified two-stage training strategy to both PredFormer and SimVP in order to improve training stability and accelerate convergence.

Pinball loss for quantile regression. For a target intensity y and a predicted q-quantile $\hat{y}^{(q)}$, the pinball loss is defined as:

$$\mathcal{L}_q(y, \hat{y}^{(q)}) = \begin{cases} q\left(y - \hat{y}^{(q)}\right), & y \geq \hat{y}^{(q)}, \\ (1-q)\left(\hat{y}^{(q)} - y\right), & y < \hat{y}^{(q)}. \end{cases} \tag{12}$$

Unlike symmetric losses such as MSE, $\mathcal{L}_q$ is *asymmetric*: under-prediction ($\hat{y}^{(q)} < y$) is penalized with weight q, whereas over-prediction ($\hat{y}^{(q)} > y$) is penalized with weight

https://doi.org/10.3390/rs18071069

$(1-q)$. Minimizing $\mathcal{L}_q$ yields a consistent estimator of the conditional q-quantile, which enables direct construction of prediction intervals from multiple quantile outputs.

Stage 1: deterministic pretraining. In the first stage, both PredFormer and SimVP are trained as deterministic forecasters targeting the median intensity by setting $q = 0.5$. In this case, the pinball loss reduces to:

$$\mathcal{L}_{0.5}(y, \hat{y}^{(0.5)}) = \frac{1}{2}\left|y - \hat{y}^{(0.5)}\right|, \tag{13}$$

which is equivalent to ℓ_1 loss up to a constant factor. Therefore, Stage 1 can be implemented using standard ℓ_1 optimization while remaining theoretically aligned with the quantile regression objective.

This pretraining stage serves two purposes: (i) it provides a stable and well-calibrated backbone representation for intensity evolution, and (ii) it acts as a warm start that significantly accelerates convergence and mitigates optimization instability when learning multiple quantiles.

Stage 2: multi-quantile fine-tuning. In the second stage, we replace the regression head with a multi-quantile head to predict K conditional quantiles $\{q_k\}_{k=1}^{K}$ at each forecast lead time. Unless otherwise specified, we use $K = 5$ quantiles $\{0.1, 0.25, 0.5, 0.75, 0.9\}$. The overall training objective averages the pinball loss over all lead times and quantile levels:

$$\mathcal{L} = \frac{1}{T_{\text{out}}K}\sum_{t=1}^{T_{\text{out}}}\sum_{k=1}^{K}\mathcal{L}_{q_k}\left(y_t, \hat{y}_t^{(q_k)}\right). \tag{14}$$

Directly training a multi-quantile head from random initialization is often sensitive to learning rates and can lead to slow or unstable calibration of prediction intervals. By contrast, the proposed two-stage scheme—Stage 1 median pretraining followed by Stage 2 quantile fine-tuning—yields more stable gradients, faster convergence, and improved reliability of the resulting prediction intervals.

4.4. Evaluation Metrics

Deterministic accuracy: The median prediction ($q = 0.5$) is treated as the point forecast and evaluated using Mean Absolute Error (MAE) and Root Mean Square Error (RMSE):

$$\text{MAE} = \frac{1}{NT_{\text{out}}}\sum_{i=1}^{N}\sum_{t=1}^{T_{\text{out}}}\left|y_{i,t} - \hat{y}_{i,t}^{(0.5)}\right|, \tag{15}$$

$$\text{RMSE} = \sqrt{\frac{1}{NT_{\text{out}}}\sum_{i=1}^{N}\sum_{t=1}^{T_{\text{out}}}\left(y_{i,t} - \hat{y}_{i,t}^{(0.5)}\right)^2}. \tag{16}$$

Probabilistic evaluation: Uncertainty quality is assessed through prediction interval calibration and sharpness. For a nominal $(1-\alpha)$ interval defined by $\left[\hat{y}_{i,t}^{(\alpha/2)}, \hat{y}_{i,t}^{(1-\alpha/2)}\right]$, we report:

- **Prediction Interval Coverage Probability (PICP):**

$$\text{PICP}_{1-\alpha} = \frac{1}{NT_{\text{out}}}\sum_{i=1}^{N}\sum_{t=1}^{T_{\text{out}}}\mathbf{1}\left(\hat{y}_{i,t}^{(\alpha/2)} \le y_{i,t} \le \hat{y}_{i,t}^{(1-\alpha/2)}\right), \tag{17}$$

which measures whether the interval is well calibrated.

https://doi.org/10.3390/rs18071069

- **Mean Prediction Interval Width (MPIW):**

$$\mathrm{MPIW}_{1-\alpha} = \frac{1}{NT_{\mathrm{out}}} \sum_{i=1}^{N} \sum_{t=1}^{T_{\mathrm{out}}} \left(\hat{y}_{i,t}^{(1-\alpha/2)} - \hat{y}_{i,t}^{(\alpha/2)} \right). \tag{18}$$

  which quantifies interval sharpness (smaller is better given adequate coverage). In this work, we primarily report 80% intervals ($\alpha = 0.2$, corresponding to the 10th and 90th percentiles) and additionally present 50% intervals (25th–75th percentiles) when analyzing uncertainty evolution across lead times.

Probabilistic evaluation metrics: Our framework produces probabilistic forecasts through direct multi-quantile regression rather than explicit parametric distributions. Accordingly, interval-based metrics such as the Prediction Interval Coverage Probability (PICP) and the Mean Prediction Interval Width (MPIW) are used to evaluate calibration and sharpness of the predicted intervals. In addition, we report the Continuous Ranked Probability Score (CRPS), a strictly proper scoring rule widely used in probabilistic meteorological forecasting. For models producing discrete quantile outputs, CRPS is approximated via numerical integration over the predicted quantiles. These complementary metrics together provide a comprehensive assessment of both deterministic accuracy and probabilistic reliability.

4.5. Implementation Details

All experiments are conducted within the OpenSTL framework, which provides standardized implementations of spatiotemporal forecasting baselines and a unified training pipeline.

Input preprocessing: Satellite images are resized to 128×128 pixels using bilinear interpolation. Unless otherwise specified, only IR and WV channels are used as visual inputs. The historical input length is $T_{\mathrm{in}} = 4$ and the prediction horizon is $T_{\mathrm{out}} = 8$. For models with scalar conditioning, historical scalar sequences are normalized using statistics computed from the training set and injected via the proposed SIFB.

PredFormer configuration: PredFormer adopts a patch-based transformer architecture with an input image size of 128×128 and two channels (IR and WV). The model uses $T_{\mathrm{in}}/T_{\mathrm{out}} = 4/8$, a patch size of 8, an embedding dimension of 256, eight attention heads, and a transformer depth of 6, consisting of one stacked temporal–spatial block.

Optimization and training strategy: All neural network models, including PredFormer, SimVP, ConvLSTM, and PredRNN, are optimized using the AdamW optimizer with a weight decay of 1×10^{-2}. The batch size is set to 8, and data loading is performed with four worker processes.

For PredFormer and SimVP, the proposed two-stage training strategy is employed. In Stage 1 (deterministic pretraining), models are trained for 15 epochs using ℓ_1 loss (equivalent to pinball loss with $q = 0.5$), with a learning rate set to 5×10^{-6} and scheduled using a One Cycle learning rate policy. In Stage 2 (multi-quantile fine-tuning), the regression head is replaced with a multi-quantile head, and the entire network (including the backbone) is fine-tuned for 35 epochs using the pinball loss with a fixed learning rate of 3×10^{-6}.

Recurrent baselines (ConvLSTM and PredRNN) are trained in a single stage for 50 epochs using ℓ_1 loss with a learning rate of 5×10^{-4}.

Scalar-only baselines. To evaluate purely scalar-based forecasting performance, we consider two scalar-only baselines. The persistence baseline (S1) extrapolates the last observed $V_{\max}$ value to all future lead times and does not require training. The regression baseline (S3) employs a simple multilayer perceptron (MLP) that maps four historical $V_{\max}$

https://doi.org/10.3390/rs18071069

values to eight future predictions. The S3 model is trained for 50 epochs using a learning rate of 5×10^{-4}.

Computing resources: All experiments are conducted on a single NVIDIA RTX 4090 GPU (24 GB memory).

5. Results

5.1. Deterministic Baselines

As shown in Table 4, the intensity scalar-only baselines, including persistence and MLP-based regression, achieve lower errors than several image-based models. This behavior is expected, as these methods directly extrapolate historical $V_{\max}$ values and therefore benefit from the strong short-term temporal autocorrelation inherent in TC intensity evolution. However, such methods rely solely on past intensity records and do not incorporate any information on storm structure or environmental evolution, which fundamentally limits their applicability beyond very short lead times and under rapidly evolving conditions.

Table 4. Deterministic forecasting performance on TCIR for 24 h (8×3 h) tropical cyclone intensity prediction. MAE and RMSE are computed from the median ($q = 0.5$) forecasts and averaged over all eight lead times. Deep learning models are reported as mean $\pm$ standard deviation across three runs with different random seeds. **Scalar-only baselines use historical $V_{\max}$ as input, whereas image-based models rely solely on IR and WV satellite imagery**.

Model	MAE (kt)	RMSE (kt)
Persistence (Scalar-only)	8.74	13.59
MLP (Scalar-only)	7.17	11.52
ConvLSTM (IR+WV)	15.94 ± 0.58	22.38 ± 0.73
PredRNN (IR+WV)	18.01 ± 0.94	25.52 ± 1.04
SimVP (IR+WV)	12.53 ± 0.62	17.41 ± 0.81
PredFormer (IR+WV)	$\mathbf{11.29 \pm 0.19}$	$\mathbf{15.71 \pm 0.12}$

Note: All metrics are evaluated on the test set. MAE and RMSE are computed using the median prediction ($q = 0.5$) and averaged over all eight forecast lead times (3–24 h). Deep learning models are evaluated with three different random seeds, and results are reported as mean $\pm$ standard deviation.

Among the image-based models that use only IR and WV satellite imagery, PredFormer achieves the lowest MAE and RMSE, outperforming both CNN-based (SimVP) and RNN-based (ConvLSTM and PredRNN) alternatives. This result suggests that the transformer-based spatiotemporal representations learned by PredFormer are more effective at capturing the evolution of TC intensity from satellite observations. In contrast, recurrent models exhibit substantially larger errors, which may be attributed to their autoregressive nature and the accumulated prediction errors over longer forecast horizons.

To further examine error growth with lead time, we report the lead-time-dependent MAE in Figure 4. The proposed TC-QFormer maintains lower errors across most lead times, and the gap becomes more evident at longer horizons, indicating improved robustness for multi-step intensity evolution forecasting.

5.2. Probabilistic Forecasting Performance

Even without scalar fusion, PredFormer already produces narrower prediction intervals than SimVP while maintaining comparable coverage, suggesting more compact and informative uncertainty estimates (Table 5). When historical intensity ($V_{\max}$) is incorporated through the proposed SIFB, both models exhibit clear improvements in deterministic accuracy and uncertainty calibration. In particular, *TC-QFormer* (ours) achieves the best overall performance on average across all lead times, featuring substantially reduced MAE

https://doi.org/10.3390/rs18071069

and RMSE, $PICP_{80}$ values close to the nominal level, and the narrowest $MPIW_{80}$ among all compared methods.

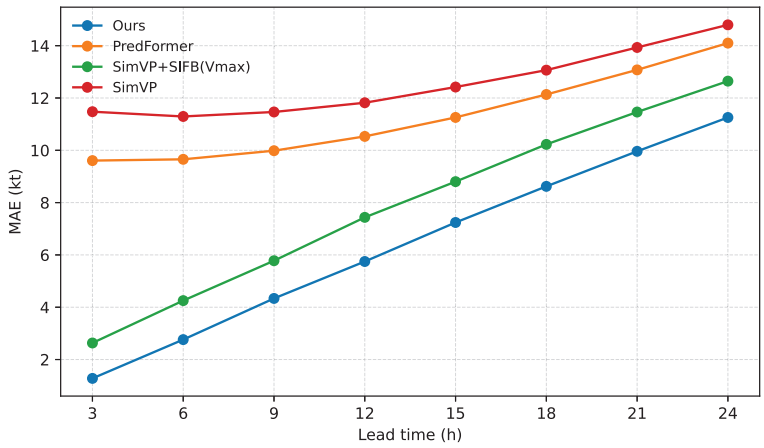


Figure 4. Lead-time-dependent deterministic forecasting accuracy (MAE) over the 24 h prediction horizon (3–24 h at 3 h intervals). The proposed TC-QFormer (ours) consistently achieves lower MAE across most lead times compared with PredFormer and SimVP-based baselines.

Table 5. Probabilistic forecasting performance on TCIR for 24 h (8×3 h) tropical cyclone intensity prediction. MAE and RMSE are computed from the median forecast ($q = 0.5$). CRPS denotes the Continuous Ranked Probability Score, approximated from discrete quantile forecasts by numerical integration. PICP and MPIW correspond to prediction intervals derived from quantile outputs. Deep learning models are reported as mean ± standard deviation across three runs with different random seeds. Image-based models use IR and WV satellite imagery as input, while models with SIFB additionally incorporate historical V_{max}. MAE, RMSE, CRPS and MPIW are measured in knots (kt).

Model	MAE	RMSE	CRPS	$PICP_{80}$	$MPIW_{80}$	$PICP_{50}$	$MPIW_{50}$
ConvLSTM	15.94 ± 0.58	22.38 ± 0.73	–	–	–	–	–
ConvLSTM + SIFB	10.87 ± 0.51	17.86 ± 0.66	–	–	–	–	–
SimVP	12.53 ± 0.62	17.41 ± 0.81	8.07 ± 0.40	0.755 ± 0.008	34.96 ± 2.46	0.451 ± 0.011	17.75 ± 1.28
SimVP + SIFB	7.90 ± 0.55	12.18 ± 0.73	5.14 ± 0.36	0.737 ± 0.007	21.57 ± 1.94	0.425 ± 0.010	10.39 ± 1.05
PredFormer	11.29 ± 0.09	15.71 ± 0.12	7.30 ± 0.06	0.715 ± 0.011	28.59 ± 0.92	0.416 ± 0.021	14.41 ± 0.67
Ours	**6.40 ± 0.05**	**10.49 ± 0.07**	**4.16 ± 0.03**	**0.772 ± 0.009**	**18.79 ± 0.63**	**0.479 ± 0.018**	**9.40 ± 0.45**

Note: $PICP_{80}$ and $MPIW_{80}$ are computed using the 10th and 90th percentiles, while $PICP_{50}$ and $MPIW_{50}$ correspond to the 25th and 75th percentiles. All metrics are evaluated on the test set and averaged over the eight forecast lead times (3–24 h). CRPS is approximated from discrete quantile forecasts by numerical integration.

Compared to SimVP-based variants, PredFormer consistently produces sharper prediction intervals under similar coverage levels, both with and without scalar fusion. This suggests that the transformer-based backbone not only improves point prediction accuracy but also facilitates more reliable and compact uncertainty quantification, which is particularly desirable for operational TC intensity forecasting.

Figure 5 shows the quantile calibration relationship averaged over lead times. The proposed method maintains empirical coverage that closely follows the nominal quantile levels, suggesting reliable uncertainty estimation.

Lead-time-dependent uncertainty characteristics are further summarized in Figure 6. Across all forecast horizons, the proposed method maintains $PICP_{80}$ values that remain closer to the nominal level while producing relatively narrower $MPIW_{80}$, suggesting more informative and sharper prediction intervals over the 24 h forecast window.

https://doi.org/10.3390/rs18071069

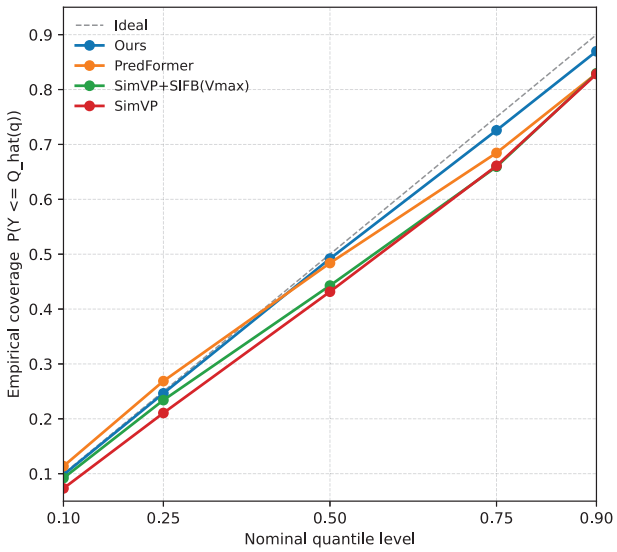
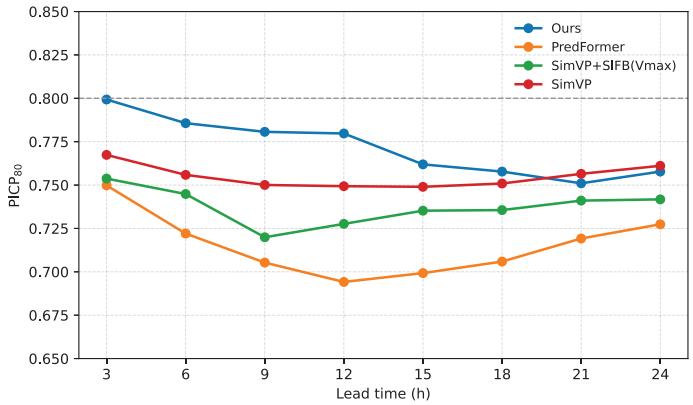
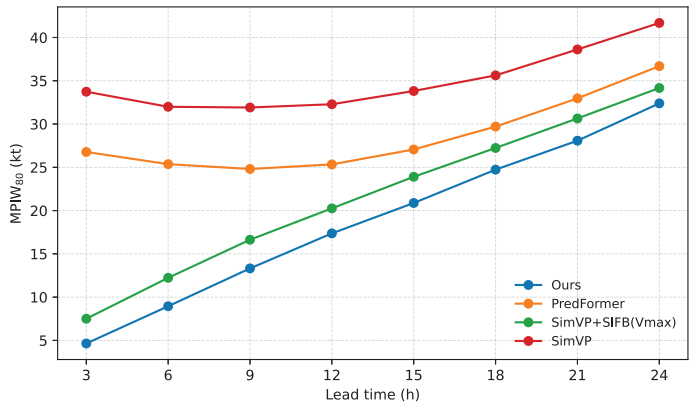


Figure 5. Quantile calibration curve averaged over forecast lead times. The horizontal axis denotes nominal quantile levels, and the vertical axis represents empirical coverage. The dashed diagonal line indicates ideal calibration ($y = x$). The proposed TC-QFormer (ours) exhibits coverage values close to the nominal levels across quantiles, indicating well-calibrated probabilistic predictions.



(**a**) Prediction Interval Coverage Probability (PICP_{80}).



(**b**) Mean Prediction Interval Width (MPIW_{80}).

Figure 6. Lead-time-dependent probabilistic evaluation metrics for 80% prediction intervals. The proposed method maintains coverage close to the nominal level while producing relatively narrow prediction intervals across forecast horizons.

https://doi.org/10.3390/rs18071069

5.3. Basin-Wise Performance Analysis

Since the TCIR dataset contains tropical cyclones from multiple ocean basins with different environmental characteristics and storm structures, we further examine whether the forecasting performance of the proposed framework varies across basins. Table 6 summarizes the basin-wise forecasting performance of TC-QFormer under the best-performing configuration (IR + WV + $V_{\max}$).

Table 6. Basin-wise forecasting performance of TC-QFormer on the test set. MAE and RMSE are computed from the median prediction ($q = 0.5$), while PICP_{80} and MPIW_{80} correspond to the 80% prediction interval. All metrics are averaged over the eight forecast lead times (3–24 h).

Basin	#Samples	MAE (kt)	RMSE (kt)	PICP_{80}	MPIW_{80} (kt)
WP	2322	7.19	11.33	0.772	21.82
EP	2340	5.97	10.28	0.786	17.60
CP	336	5.26	7.94	0.820	18.77
NA	1478	5.13	8.29	0.771	15.39
IO	324	6.57	10.50	0.712	16.69
SH	1626	7.25	11.71	0.754	19.71
Overall	8426	6.40	10.49	0.772	18.79

Overall, the forecasting errors remain within a relatively narrow range across basins, suggesting that the proposed framework maintains reasonably consistent forecasting skill under different regional storm characteristics. Slightly larger errors are observed in the western Pacific and Southern Hemisphere basins, which may reflect the greater diversity of storm structures and environmental conditions in these regions.

The probabilistic metrics also exhibit similar behavior across basins. In particular, the PICP_{80} values remain close to the nominal coverage level, while MPIW_{80} stays within a comparable range across regions, indicating that the proposed model provides stable uncertainty estimation under varying basin conditions.

5.4. Ablation on Scalar–Image Fusion

An ablation study analyzing the impact of different scalar configurations within the SIFB framework is presented in Table 7. Incorporating only historical $V_{\max}$ leads to a pronounced improvement over the image-only baseline across all metrics, including MAE, RMSE, and interval sharpness. This result indicates that recent intensity history provides the most informative scalar cue for short-term TC intensity forecasting and can be effectively integrated with satellite imagery through the proposed fusion mechanism.

Table 7. Ablation study on the Scalar–Image Fusion Block (SIFB) for 24 h (8×3 h) tropical cyclone intensity forecasting. MAE and RMSE are computed from the median forecast ($q = 0.5$), while PICP and MPIW correspond to prediction intervals derived from quantile outputs. All metrics are evaluated on the test set and averaged over all eight lead times.

Backbone	Scalar Set	MAE (kt)	RMSE (kt)	PICP_{80}	MPIW_{80} (kt)	PICP_{50}	MPIW_{50} (kt)
PredFormer	None	11.29	15.71	0.715	28.59	0.416	14.41
PredFormer	$V_{\max}$ only	**6.40**	10.49	**0.772**	18.79	**0.479**	9.40
PredFormer	MSLP only	8.72	12.41	0.678	21.03	0.395	10.53
PredFormer	$V_{\max}$ + MSLP	6.56	**10.45**	0.751	**17.56**	0.437	**8.41**
PredFormer	5 Scalars	7.05	10.96	0.730	18.52	0.441	9.26
SimVP	$V_{\max}$ only	7.90	12.18	0.737	21.57	0.425	10.39

Note: PICP_{80} and MPIW_{80} are computed using the 10th and 90th percentiles. PICP_{50} and MPIW_{50} correspond to the interquartile range (25th–75th percentiles). Smaller MPIW indicates sharper prediction intervals given comparable coverage.

https://doi.org/10.3390/rs18071069

When additional scalar variables are introduced, performance slightly decreases compared with the $V_{\max}$-only configuration, although it remains clearly superior to the no-scalar baseline. A plausible explanation is that historical $V_{\max}$ values are strongly correlated with short-term future intensity, making them the most informative scalar predictor in this setting. In contrast, additional variables such as storm size indicators (e.g., $R35_{4qAVG}$) or geographic location (latitude and longitude) are only indirectly related to short-term intensity change, and their statistical relationships with near-term intensity evolution are comparatively weaker. Introducing these variables may therefore introduce additional noise that partially dilutes the strong predictive signal provided by $V_{\max}$ alone. It is also noteworthy that using MSLP alone still yields a substantial improvement over the no-scalar baseline, and performance remains stable when MSLP is combined with $V_{\max}$. This behavior is consistent with the well-known physical relationship between maximum wind speed and minimum sea-level pressure in tropical cyclones, where stronger winds are generally associated with lower central pressure.

Another factor may lie in the simplicity of the current fusion mechanism. The SIFB is intentionally designed as a lightweight conditioning module to maintain computational efficiency, and it therefore tends to capture dominant low-order statistical relationships between scalar inputs and the prediction target. While this design effectively exploits the strong predictive signal provided by $V_{\max}$, more sophisticated fusion strategies that explicitly account for the heterogeneous physical roles and uncertainties of different environmental variables may further improve performance when multiple scalar predictors are introduced.

Across all scalar configurations, the proposed SIFB consistently improves both deterministic and probabilistic performance relative to image-only models. A direct comparison between PredFormer and SimVP under the identical scalar configuration further highlights the advantage of the transformer-based backbone, as PredFormer achieves lower errors and sharper prediction intervals when conditioned on identical scalar inputs.

5.5. Overall Discussion of Results

Overall, the experimental results lead to three main conclusions. First, satellite imagery provides essential information for TC intensity forecasting beyond simple extrapolation of historical intensity, particularly by capturing the evolving storm structure absent in scalar-only approaches. Second, transformer-based spatiotemporal modeling with PredFormer yields superior performance over CNN- and RNN-based baselines for both deterministic and probabilistic prediction, highlighting its effectiveness in learning structure–intensity relationships from satellite observations. Third, the proposed Scalar–Image Fusion Block effectively leverages historical intensity information—particularly $V_{\max}$—to further improve forecasting accuracy and uncertainty quantification. In addition, basin-wise analysis suggests that the proposed framework maintains reasonably consistent forecasting performance across different ocean basins.

6. Case Studies

To complement the quantitative evaluation in Section 5, we conduct qualitative and semi-quantitative analyses under different TC intensity evolution regimes. Particular attention is given to rapid intensification (RI), which is widely recognized as one of the most challenging scenarios in operational TC forecasting due to its strong nonlinearity and limited predictability. In this study, RI is defined as $\Delta V_{\max,24h} = V_{\max}(t{+}24) - V_{\max}(t) \geq 30$ kt. We first examine two representative RI cases with different intensification magnitudes, followed by an aggregate comparison between RI and non-RI regimes.

https://doi.org/10.3390/rs18071069

6.1. Rapid Intensification Scenarios

6.1.1. Mild RI Case (30–40 kt Increase)

Figure 7 shows that all models capture the overall upward trend in intensity; however, notable differences arise in the timing and amplitude of intensification. The proposed framework more closely follows the observed growth rate, resulting in smaller cumulative deviation over the 24 h forecast horizon. Although some underestimation of the peak intensity remains, the amplitude error is reduced relative to both baseline models.

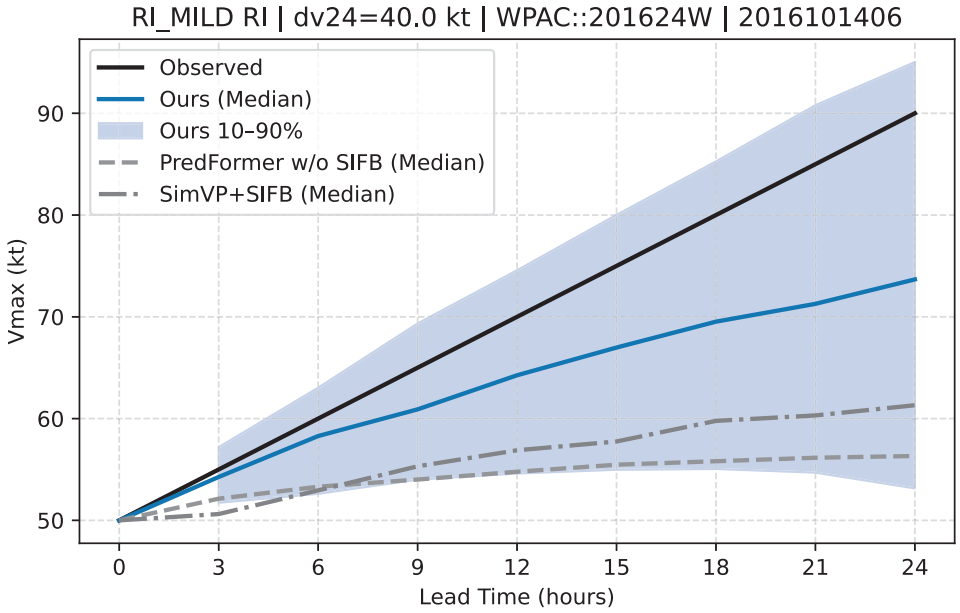


Figure 7. Mild rapid intensification (RI) case from the test set ($\Delta V_{\max,24h} = 40$ kt; WPAC 2016). Black line: observed best-track $V_{\max}$. Blue solid line: proposed model median forecast ($q = 0.5$). Light blue shaded region: 10–90% prediction interval. Gray dashed line: PredFormer without SIFB (median). Dark gray dash-dotted line: SimVP+SIFB (median).

From a probabilistic perspective, the predicted 10–90% interval expands during the rapid growth phase, reflecting increased forecast uncertainty under dynamically evolving conditions. At longer lead times, the observed intensity trajectory largely remains within the predicted interval, suggesting reasonable probabilistic calibration for this moderate RI scenario.

6.1.2. Strong RI Case (45–55 kt Increase)

As illustrated in Figure 8, all models exhibit increased difficulty in capturing the peak intensity, highlighting the challenge of forecasting extreme intensification events. The proposed framework follows the overall intensification trend more closely than the baseline models, although underestimation of the maximum intensity persists. PredFormer without scalar fusion shows larger amplitude deviations and noticeable short-term oscillations, which may result from relying solely on visual features without explicit scalar conditioning on recent intensity history. In contrast, SimVP+SIFB produces smoother trajectories but exhibits persistent underestimation throughout the forecast horizon.

The prediction interval generated by the proposed model expands substantially during the most rapidly intensifying phase. Despite the challenging dynamics, while some under-coverage remains, the interval still captures a substantial portion of the observed trajectory, reflecting heightened predictive uncertainty under strong RI conditions.

https://doi.org/10.3390/rs18071069

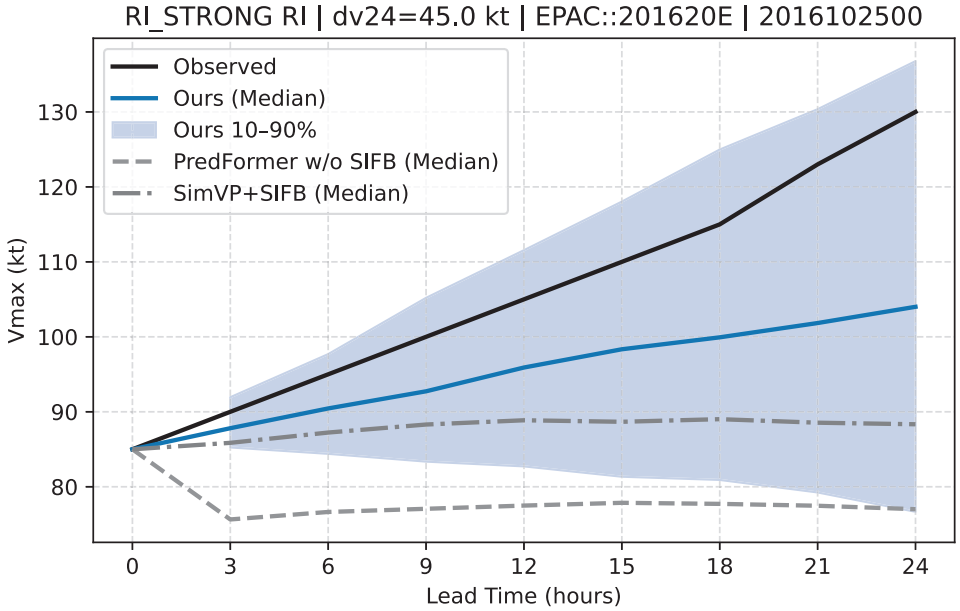


Figure 8. Strong rapid intensification (RI) case from the test set ($\Delta V_{\max,24h} = 45$ kt; EPAC 2016). Black line: observed best-track $V_{\max}$. Blue solid line: proposed model median forecast ($q = 0.5$). Light blue shaded region: 10–90% prediction interval. Gray dashed line: PredFormer without SIFB (median). Dark gray dash-dotted line: SimVP+SIFB (median).

6.2. Comparison with Baseline Models

Consistent patterns emerge across both RI scenarios. Incorporating SIFB improves the tracking of intensification rate and reduces cumulative deviation relative to the backbone-only PredFormer. Meanwhile, the transformer-based backbone provides improved temporal stability under rapidly evolving conditions compared with SimVP. These case-level results suggest that combining global spatiotemporal modeling with scalar conditioning enhances model robustness during periods of rapid structural transitions.

6.3. RI Versus Non-RI Error Analysis

To quantitatively assess forecast difficulty under rapid intensification, we partition the test dataset into RI and non-RI subsets using the 24 h definition described above. Among 8426 test samples, 689 (8.2%) satisfy the RI criterion, while the remaining 7737 samples correspond to non-RI conditions (Table 8). Forecast performance degrades substantially for all models under RI conditions. For the proposed framework, MAE increases from 5.72 kt in non-RI cases to 14.08 kt in RI cases, reflecting the heightened intrinsic variability and reduced predictability associated with rapid intensity growth. Nevertheless, the proposed model maintains consistent advantages over both baseline methods across both regimes. Under RI conditions, it reduces MAE by 37.1% relative to PredFormer without SIFB and by 25.5% relative to SimVP+SIFB.

From an uncertainty perspective, prediction interval coverage decreases for all models during RI events. For example, PICP_{80} drops to 0.537 for the proposed model, indicating that extreme intensification remains challenging to fully capture. However, despite this degradation, the proposed framework maintains relatively higher PICP_{80} with narrower MPIW_{80} compared with baseline methods in both RI and non-RI subsets, indicating a more favorable coverage–sharpness trade-off in relative terms.

https://doi.org/10.3390/rs18071069

Table 8. Performance under rapid intensification (RI) and non-RI conditions on the test set. RI is defined as $\Delta V_{\max,24h} \geq 30$ kt. MAE and RMSE are computed from the median forecast; PICP80 and MPIW80 correspond to the 10–90% prediction interval.

Model	Regime	MAE (kt)	RMSE (kt)	PICP80	MPIW80 (kt)
Ours	RI	14.08	19.04	0.537	24.31
	Non-RI	5.72	9.35	0.793	18.30
PredFormer w/o SIFB	RI	22.37	27.91	0.444	34.01
	Non-RI	10.31	14.12	0.740	28.11
SimVP+SIFB	RI	18.90	24.34	0.397	26.28
	Non-RI	6.93	10.43	0.768	21.15

6.4. Grad-CAM Visualization of Model Attention

To further examine the spatial cues utilized by the proposed model, Grad-CAM visualizations are provided for one representative non-RI case and one RI case, as shown in Figure 9. The visualization highlights the regions in the input satellite imagery that contribute most strongly to the predicted intensity change.

For the non-RI example (Figure 9, left), the activation is mainly distributed over several outer convective clusters and peripheral cloud-band regions, while the storm center exhibits relatively weak response. This suggests that for storms with relatively stable short-term intensity evolution, the model relies more on the spatial distribution of surrounding convection rather than a strongly organized inner-core structure.

In contrast, the RI example (Figure 9, right) shows that the Grad-CAM activation is concentrated closer to the storm center and nearby deep convective regions. The highlighted areas align with the more organized inner-core cloud structures visible in the IR and WV imagery, indicating that the model places greater emphasis on central convective organization when predicting rapid intensification.

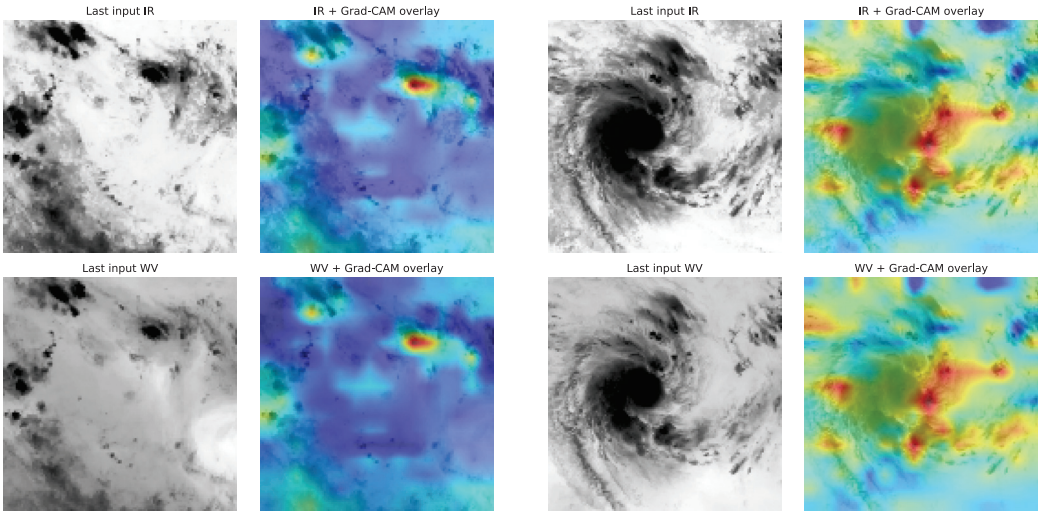


Figure 9. Grad-CAM visualization of representative non-RI and RI cases from the test set. (**Left**): non-RI case (ATLN 2016), where model attention is mainly distributed over outer convective clusters and peripheral cloud-band regions. (**Right**): RI case (WPAC 2016), where the activation is concentrated closer to the storm center and nearby deep convective regions. These patterns indicate that the proposed model adaptively focuses on different spatial structures under different intensity-evolution regimes.

Overall, these visualizations suggest that the model adaptively shifts its spatial attention depending on the storm evolution regime. Outer convective structures appear more

https://doi.org/10.3390/rs18071069

relevant for non-RI storms, whereas inner-core convective organization becomes more important during rapid intensification.

7. Discussion

Extending tropical cyclone intensity forecasting from deterministic point prediction to interval-based prediction provides operationally valuable uncertainty information. In practice, forecasters are concerned not only with the most likely intensity estimate but also with the plausible range of future intensity evolution, which is critical for risk assessment and warning decisions. In the proposed framework, predictive intervals are obtained through simultaneous estimation of multiple conditional quantiles. For example, the interval defined by the 0.1 and 0.9 quantiles corresponds to an empirical 80% prediction interval, while the interval between the 0.25 and 0.75 quantiles represents a 50% prediction interval. The reliability of these intervals is evaluated using coverage- and sharpness-related metrics such as PICP, MPIW, and calibration analysis. Compared with prior transformer-based TC forecasting studies that focus mainly on deterministic prediction, this work explicitly formulates TC intensity forecasting as an interval prediction task. To our knowledge, relatively few studies jointly model tropical cyclone satellite imagery together with scalar storm information while directly producing intensity prediction intervals.

Recent studies have explored probabilistic tropical cyclone prediction using generative approaches such as diffusion models. These methods approximate predictive distributions through iterative sampling and therefore provide strong flexibility in representing complex probabilistic structures. However, diffusion-based approaches typically require substantially higher training cost and more complex tuning procedures, particularly when adapting to heterogeneous observational inputs. In contrast, the proposed framework estimates target quantiles in a single forward pass without additional sampling steps, resulting in a computationally efficient pipeline that is also flexible for integrating different types of operational input data.

The results indicate that combining spatiotemporal visual modeling with scalar conditioning is an effective strategy for probabilistic TC intensity forecasting. Using a non-autoregressive transformer backbone, the framework captures global spatiotemporal context from satellite imagery while avoiding the error accumulation that commonly affects recurrent models at longer lead times. The SIFB provides a simple and interpretable mechanism for incorporating historical intensity information, enabling the model to adaptively modulate visual features under different intensity evolution regimes. Case-level analyses suggest that the framework exploits physically meaningful structural cues in satellite imagery—such as inner-core organization and surrounding cloud evolution—rather than relying solely on extrapolation of past intensity values. This behavior is particularly evident in rapid intensification (RI) cases, where scalar conditioning helps stabilize intensity growth signals while the spatiotemporal backbone captures evolving structural patterns. As expected, prediction errors are higher for RI cases than for non-RI cases across all models, reflecting the intrinsic nonlinearity and limited predictability of rapid intensity changes as well as the relatively small number of RI samples available for training. Nevertheless, the proposed framework maintains consistent performance advantages over baseline methods in both regimes, and prediction intervals appropriately expand during RI periods, reflecting increased predictive uncertainty.

Several limitations of the present study should be acknowledged. First, the current work focuses on short-term (24 h) intensity forecasting. As forecast horizons increase, prediction errors naturally grow due to the nonlinear and chaotic nature of tropical cyclone intensity evolution. For relatively weak storms or systems with stable intensity trends, structural information from satellite imagery may remain useful over moderate lead times.

https://doi.org/10.3390/rs18071069

However, for rapidly evolving systems—particularly those undergoing rapid intensification or structural transitions—forecast uncertainty increases substantially, and accurately capturing turning points in intensity evolution becomes more difficult. Extending the framework to longer lead times (e.g., 48–120 h) will likely require incorporating additional environmental predictors together with potential physical constraints or postprocessing strategies. Notably, the proposed framework was designed with extensibility in mind: the scalar–image fusion mechanism allows additional environmental variables to be integrated naturally, providing a flexible foundation for future extensions toward longer-range intensity forecasting. Second, the current implementation relies only on IR and WV satellite imagery and does not explicitly incorporate additional observational modalities or environmental predictors (e.g., ocean heat content or vertical wind shear), which may provide complementary information for challenging scenarios and longer forecast horizons. Third, uncertainty is represented using a finite set of quantiles, which provides interpretable prediction intervals but does not produce a fully continuous predictive distribution. Nevertheless, the proposed framework demonstrates a favorable balance among forecasting accuracy, uncertainty interpretability, and computational efficiency.

8. Conclusions

This study proposes an interval-based probabilistic framework for 24 h TC intensity forecasting that combines transformer-based spatiotemporal modeling with scalar-conditioned feature fusion and multi-quantile regression. Using the TCIR dataset, we demonstrate that the proposed approach achieves improved deterministic accuracy while producing well-calibrated and operationally interpretable prediction intervals. The results show that transformer-based spatiotemporal modeling with PredFormer consistently outperforms CNN- and RNN-based baselines for both point prediction and uncertainty quantification. Incorporating historical intensity through SIFB further improves forecasting performance, particularly by improving the representation of intensity evolution under rapidly changing conditions. Overall, these findings suggest that conditional spatiotemporal modeling offers a practical and effective direction for probabilistic TC intensity forecasting, bridging the gap between purely deterministic predictions and more complex probabilistic approaches. The proposed framework provides a transparent and computationally efficient solution that is well suited for short-term operational forecasting and risk-informed decision support.

Author Contributions: Conceptualization, T.G., H.Z. and S.P.; methodology, T.G. and H.Z.; software, T.G.; validation, T.G. and H.Z.; formal analysis, T.G. and H.Z.; investigation, T.G.; data curation, T.G.; writing—original draft preparation, T.G.; writing—review and editing, H.Z., T.S. and S.P.; visualization, T.G.; supervision, S.P.; project administration, S.P. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No. XDA0500202).

Data Availability Statement: The TCIR dataset used in this study is publicly available at: https://github.com/BoyoChen/TCIR (accessed on 29 March 2026) and the associated HDF5 files can be downloaded from https://drive.google.com/drive/folders/19Jo9iLi2b5qFoPan78SyojBxDHr-egJw (accessed on 29 March 2026). The OpenSTL training framework is publicly available at: https://github.com/chengtan9907/OpenSTL (accessed on 29 March 2026). The PredFormer backbone implementation is available at: https://github.com/yyyujintang/PredFormer (accessed on 29 March 2026). The implementation of the proposed TC-QFormer framework will be made publicly available upon publication.

Conflicts of Interest: The authors declare no conflicts of interest.

https://doi.org/10.3390/rs18071069

References

1. Emanuel, K. Increasing destructiveness of tropical cyclones over the past 30 years. *Nature* **2005**, *436*, 686–688. [CrossRef] [PubMed]
2. Pielke, R.A.J.; Gratz, J.; Landsea, C.W.; Collins, D.; Saunders, M.A.; Musulin, R. Normalized hurricane damage in the United States: 1900–2005. *Nat. Hazards Rev.* **2008**, *9*, 29–42. [CrossRef]
3. Knutson, T.R.; McBride, J.L.; Chan, J.; Emanuel, K.; Holland, G.; Landsea, C.; Held, I.; Kossin, J.P.; Srivastava, A.K.; Sugi, M. Tropical cyclones and climate change. *Nat. Geosci.* **2010**, *3*, 157–163. [CrossRef]
4. Gall, R.; Franklin, J.; Marks, F.; Rappaport, E.N.; Toepfer, F. The Hurricane Forecast Improvement Project. *Bull. Am. Meteorol. Soc.* **2013**, *94*, 329–343. [CrossRef]
5. Emanuel, K.A. 100 Years of Progress in Tropical Cyclone Research. *Meteorol. Monogr.* **2018**, *59*, 15.1–15.68. [CrossRef]
6. Emanuel, K.A. An air-sea interaction theory for tropical cyclones. Part I: Steady-state maintenance. *J. Atmos. Sci.* **1986**, *43*, 585–604. [CrossRef]
7. Montgomery, M.T.; Smith, R.K. Recent developments in the fluid dynamics of tropical cyclones. *Annu. Rev. Fluid Mech.* **2017**, *49*, 541–574. [CrossRef]
8. Rogers, R.F.; Reasor, P.D.; Lorsolo, S. Airborne Doppler observations of the inner-core structural differences between intensifying and steady-state tropical cyclones. *Mon. Weather Rev.* **2013**, *141*, 2970–2991. [CrossRef]
9. DeMaria, M.; Kaplan, J. A Statistical Hurricane Intensity Prediction Scheme (SHIPS) for the Atlantic Basin. *Weather Forecast.* **1994**, *9*, 209–220. [CrossRef]
10. DeMaria, M.; Mainelli, M.; Shay, L.K.; Knaff, J.A.; Kaplan, J. Further Improvements to the Statistical Hurricane Intensity Prediction Scheme (SHIPS). *Weather Forecast.* **2005**, *20*, 531–543. [CrossRef]
11. Dvorak, V.F. *Tropical Cyclone Intensity Analysis and Forecasting from Satellite Imagery*; Technical Report NOAA Technical Memorandum NESS 45; National Oceanic and Atmospheric Administration (NOAA): Silver Spring, MD, USA, 1975.
12. Velden, C.; Harper, B.; Wells, F.; Beven, J.; Zehr, R.; Olander, T.; Mayfield, M.; Guard, C.; Lander, M.; Edson, R. The Dvorak tropical cyclone intensity estimation technique: A satellite-based method that has endured for over 30 years. *Bull. Am. Meteorol. Soc.* **2006**, *87*, 1195–1210. [CrossRef]
13. Kaplan, J.; DeMaria, M.; Knaff, J.A. A Revised Tropical Cyclone Rapid Intensification Index for the Atlantic and Eastern North Pacific Basins. *Weather Forecast.* **2010**, *25*, 220–241. [CrossRef]
14. Cangialosi, J.P.; Blake, E.S.; DeMaria, M.; Latto, A.S.; Pasch, R.J.; Rappaport, E.N. *Recent Progress in Tropical Cyclone Intensity Forecasting at the National Hurricane Center*; Technical report; National Hurricane Center: Miami, FL, USA, 2020.
15. Pradhan, R.; Aygün, R.S.; Maskey, M.; Ramachandran, R.; Cecil, D.J. Tropical Cyclone Intensity Estimation Using a Deep Convolutional Neural Network. *IEEE Trans. Image Process.* **2018**, *27*, 692–702. [CrossRef] [PubMed]
16. Chen, B.F.; Chen, B.; Lin, H.T.; Elsberry, R.L. Estimating Tropical Cyclone Intensity by Satellite Imagery Utilizing Convolutional Neural Networks. *Weather Forecast.* **2019**, *34*, 447–465. [CrossRef]
17. Shi, X.; Chen, Z.; Wang, H.; Yeung, D.Y.; Wong, W.K.; Woo, W.c. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. In Proceedings of the Advances in Neural Information Processing Systems 28 (NeurIPS 2015), Montreal, QC, Canada, 7–12 December 2015; pp. 802–810.
18. Wang, Y.; Long, M.; Wang, J.; Gao, Z.; Yu, P.S. PredRNN: Recurrent Neural Networks for Predictive Learning using Spatiotemporal LSTMs. In Proceedings of the Advances in Neural Information Processing Systems 30 (NeurIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 879–888.
19. Gao, Z.; Tan, C.; Wu, L.; Li, S.Z. SimVP: Simpler Yet Better Video Prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 3170–3180. [CrossRef]
20. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. In Proceedings of the Advances in Neural Information Processing Systems 30 (NeurIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
21. Lim, B.; Zohren, S. Time-series forecasting with deep learning: A survey. *Philos. Trans. R. Soc. A* **2021**, *379*, 20200209. [CrossRef]
22. Zhou, H.; Zhang, S.; Peng, J.; Zhang, S.; Li, J.; Xiong, H.; Zhang, W. Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 11106–11115. [CrossRef]
23. Wu, H.; Xu, J.; Wang, J.; Long, M. Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting. In Proceedings of the Advances in Neural Information Processing Systems 34 (NeurIPS 2021), Virtual, 6–14 December 2021; pp. 22419–22430.
24. Wu, H.; Hu, T.; Liu, Y.; Zhou, H.; Wang, J.; Long, M. TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis. In Proceedings of the International Conference on Learning Representations (ICLR), Kigali, Rwanda, 1–5 May 2023.
25. Tang, Y.; Qi, L.; Xie, F.; Li, X.; Ma, C.; Yang, M.H. Video Prediction Transformers without Recurrence or Convolution. *arXiv* **2024**, arXiv:2410.04733. [CrossRef]

https://doi.org/10.3390/rs18071069

26. Jiang, W.; Zhang, D.; Hu, G.; Wu, T.; Liu, L.; Xiao, Y.; Duan, Z. Transformer-based tropical cyclone track and intensity forecasting. *J. Wind Eng. Ind. Aerodyn.* **2023**, *238*, 105440. [CrossRef]
27. Gneiting, T.; Raftery, A.E. Strictly Proper Scoring Rules, Prediction, and Estimation. *J. Am. Stat. Assoc.* **2007**, *102*, 359–378. [CrossRef]
28. Gneiting, T.; Katzfuss, M. Probabilistic forecasting. *Annu. Rev. Stat. Its Appl.* **2014**, *1*, 125–151. [CrossRef]
29. Roulston, M.S.; Smith, L.A. Combining dynamical and statistical ensembles. *Tellus A Dyn. Meteorol. Oceanogr.* **2003**, *55*, 16–30. [CrossRef]
30. Pinson, P. Very-short-term probabilistic forecasting of wind power with generalized logit–normal distributions. *J. R. Stat. Soc. Ser. C (Appl. Stat.)* **2012**, *61*, 555–576. [CrossRef]
31. Koenker, R.; Bassett, G. Regression Quantiles. *Econometrica* **1978**, *46*, 33–50. [CrossRef]
32. Koenker, R. *Quantile Regression*; Number 38 in Econometric Society Monographs; Cambridge University Press: Cambridge, UK, 2005. [CrossRef]
33. Özkan, M.B.; Güvençir, U.; Küçük, D.; Seçen, A.Ü.; Buhan, S.; Demirci, T.; Bestil, A.; Er, C.; Karagöz, P. Probabilistic Wind Power Forecasting by Using Quantile Regression Analysis. In *Proceedings of the Data Analytics for Renewable Energy Integration: Informing the Generation and Distribution of Renewable Energy; Lecture Notes in Computer Science, Skopje, Macedonia, 22 September 2017*; Springer: Cham, Switzerland, 2017; Volume 10691, pp. 72–82. [CrossRef]
34. Hatalis, K.; Lamadrid, A.J.; Scheinberg, K.; Kishore, S. Smooth Pinball Neural Network for Probabilistic Forecasting of Wind Power. *arXiv* **2017**, arXiv:1710.01720. [CrossRef]
35. Xu, C.; Sun, Y.; Du, A.; Gao, D.c. Quantile regression based probabilistic forecasting of renewable energy generation and building electrical load: A state-of-the-art review. *J. Build. Eng.* **2023**, *79*, 107772. [CrossRef]
36. Goerss, J.S.; Sampson, C.R. Prediction of Consensus Tropical Cyclone Intensity Forecast Error. *Weather Forecast.* **2014**, *29*, 750–762. [CrossRef]
37. Luo, J.J.; Yang, P.; Meng, F. Probabilistic Forecasting Model for Tropical Cyclone Intensity Based on Diffusion Model. *Remote Sens.* **2025**, *17*, 3600. [CrossRef]
38. Chen, B.; Chen, B.F.; Lin, H.T. Rotation-blended CNNs on a New Open Dataset for Tropical Cyclone Image-to-intensity Regression. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, UK, 19–23 August 2018; pp. 90–99. [CrossRef]
39. Knapp, K.R.; Kruk, M.C.; Levinson, D.H.; Diamond, H.J.; Neumann, C.J. The International Best Track Archive for Climate Stewardship (IBTrACS): Unifying tropical cyclone best track data. *Bull. Am. Meteorol. Soc.* **2010**, *91*, 363–376. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

https://doi.org/10.3390/rs18071069

remote sensing

MDPI

Article

TDA-DARKNet: A Deep Learning Model Based on Dual-Polarization Radar Data for Tornado Detection

Guoxiu Zhang [1,2], Qiangyu Zeng [1,3,4], Fugui Zhang [1,2,*], Hao Wang [1,2] and Tiantian Yu [1,2]

[1] College of Electronic Engineering, Chengdu University of Information Technology, Chengdu 610225, China; 3230301015@stu.cuit.edu.cn (G.Z.); zqy@cuit.edu.cn (Q.Z.); wh@cuit.edu.cn (H.W.); yutiantian@cuit.edu.cn (T.Y.)
[2] Key Laboratory of Atmosphere Sounding, China Meteorological Administration, Chengdu 610225, China
[3] Key Laboratory of South China Sea Meteorological Disaster Prevention and Mitigation of Hainan Province, Haikou 570203, China
[4] China Meteorological Administration Tornado Key Laboratory, Beijing 100871, China
* Correspondence: zfg@cuit.edu.cn

Highlights

What are the main findings?

- We propose a novel tornado identification model, TDA-DARKNet, which combines convolutional neural networks with channel and spatial attention mechanisms to strengthen responses to key variables and spatial regions, while incorporating a Kolmogorov–Arnold Network (KAN) to enhance nonlinear representation capability.
- The model enhances tornado detection capability, particularly for weak tornado events.

What are the implications of the main findings?

- Cross-regional evaluation using multiple independent tornado events indicates that the model maintains stable detection performance and good regional generalization.
- The results suggest that the transition from "handcrafted feature-driven" approaches to "deep representation-driven" frameworks constitutes a key technological pathway for improving tornado detection probability and extending warning lead time.

Abstract

Tornado is a localized, small-scale severe convective weather phenomenon characterized by extreme destructiveness. Tornado detecting and warning mainly rely on Doppler weather radar, which identifies and tracks tornadoes by recognizing the tornado vortex signature and supercells in radar data. Artificial intelligence technology has been applied to tornado recognition in recent years. However, existing monitoring methods, especially those using unsupervised learning algorithms, still have limited recognition accuracy and timely warning, and usually struggle to strike a balance between detection accuracy and false alarm rate. A novel tornado detection algorithm TDA-DARKNet has been proposed to address the aforementioned issues. The algorithm integrates a dual attention mechanism, dense residual connections, and Kolmogorov–Arnold network (KAN). A tornado dataset suitable for deep learning has been formed, which utilizes features including radial velocity, reflectivity, velocity spectrum width, differential reflectivity, and correlation coefficient in radar data. The TDA-DARKNet algorithm was trained and tested using the tornado dataset, and evaluated in tornado cases. The experimental results show that TDA-DARKNet improves the detection probability and extends the lead time to a maximum of 42 min in strong tornado situations, while achieving 97.11% accuracy, 95.08% precision, indicating strong overall identification performance. In addition, by directly leveraging radar-based data for tornado identification, the algorithm eliminates the need for manual feature engineering, simplifies data processing, reduces complexity, and further enhances detection effectiveness.

Keywords: tornado detection; dual-polarization radar; attention mechanism; denseblock; Kolmogorov–Arnold Network (KAN)

1. Introduction

Tornadoes are small-scale weather phenomena with strong destructive potential, characterized by limited spatial extent and short lifetimes. Accurate prediction and early warning of tornadoes remain highly challenging, constituting a longstanding frontier in atmospheric science. According to statistical analyses of tornadoes in China, tornadoes typically exhibit wind speeds of 100–140 $m \cdot s^{-1}$, diameters of less than 2 km, and path lengths ranging from 0 to 25 km [1]. Tornadoes are generally classified into two categories: mesocyclone tornadoes and non-mesocyclone tornadoes, with the latter typically being weaker than the former [2,3]. Tornadoes predominantly occur in the flat terrains of the eastern and certain central regions in China. Jiangsu Province, especially its northern part, exhibits the highest frequency of tornado occurrences nationwide. The occurrence of tornadoes is closely linked to atmospheric conditions, particularly for major tornado outbreaks, which are typically accompanied by specific atmospheric anomalies and wind field conditions. Ćwik et al. [4] analyzed major tornado outbreaks in the United States during May by applying Maximum Covariance Analysis (MCA) to 91 significant tornado events from 1950 to 2019. Their findings provide new insights for understanding tornado outbreak characteristics and early warning in the context of climate change. The mechanisms of tornado formation remain incompletely understood; however, several theoretical models have been proposed to explain this process. Artekha [5], based on the Schwarzschild criterion for convective instability, analyzed the tornado velocity field structure through both analytical and numerical solutions, incorporating the stream function, azimuthal velocity components, and density change equations in cylindrical coordinates. These findings provide theoretical support for improving tornado prediction accuracy.

Weather radar is a crucial tool for tornado detection. Early tornado signature recognition algorithms primarily relied on dynamical features such as radar radial velocity for identification. The National Severe Storms Laboratory (NSSL) used pulse-Doppler radar to detect tornadoes and discovered a distinctive tornado vortex signature (TVS). Zrnić et al. [6] proposed the Mesocyclone Detection Algorithm (MDA), which identifies mesocyclones through the evaluation of spatial scale and azimuthal shear of radial velocity. MDA substantially enhanced the warning capability for tornadoes associated with supercell storms. In 1989, a new tornado detection algorithm (TDA-TVS) was developed to address the low detection rate associated with threshold-based identification of the TVS [7]. TDA-TVS successfully expanded tornado identification from supercell tornadoes to non-supercell tornadoes. It should be noted that TVS detection is strongly dependent on radar characteristics, scanning strategies, and the distance from the radar, which can significantly influence its detection capability. Steven [8] found that the integration of radar reflectivity with near-storm environmental parameters derived from numerical models significantly improves the automated classification of convective storms, particularly in identifying rotating storms and assessing their potential for severe weather. However, the rotational signals of many weak tornadoes are confined to the lower elevation levels of the radar, resulting in a relatively low detection rate.

In the 21st century, substantial progress in tornado detection has been achieved with the updating from Doppler radar to dual-polarization radar systems. Dual-polarization radar integrates fundamental Doppler parameters with polarimetric variables, enabling more comprehensive detection and analysis of meteorological phenomena.

https://doi.org/10.3390/rs18081124

Ryzhkov et al. [9] conducted a detailed investigation of three tornado events and found that tornado debris in the hook echo region was associated with low correlation coefficient ρ_{HV} (below 0.8) and differential reflectivity Z_{DR} (less than 0.5 dB). Based on these findings, a tornado detection algorithm utilizing tornado debris signatures (TDSs) was developed. Bluestein et al. [10] further pointed out that, in the identification of TDSs, the correlation coefficient is more effective in tornado detection algorithms due to its insensitivity to signal attenuation. Kumjian [11] identified distinct features such as the Z_{DR} arc, Z_{DR} column, and K_{DP} column based on observations from C-band polarimetric radar. Recognizing that mixed precipitation-debris scenarios may not produce extremely low correlation coefficient values, Ryzhkov [12] revised the TDS criteria by relaxing the correlation coefficient threshold and introducing reflectivity and azimuthal shear constraints, thereby improving operational robustness. Zhang et al. [13] conducted the observation and documentation of a waterspout over the Pearl River estuary using a dual-polarization phased array radar (dual-PAR) network in 2020. This study confirmed that tornado occurrence is associated with abnormally low values of the ρ_{HV}, consistent with known polarimetric signatures. Analysis of the prior studies reveals that TDS typically appears after tornado touchdown, which limits its effectiveness for early warning. However, it serves as a reliable indicator for confirming that a tornado has occurred.

By integrating effective features from traditional identification methods and introducing novel ones, artificial intelligence has made substantial advancements in tornado detection and warning, greatly enhancing the efficiency of both identification and early warning. Wang [14] innovatively proposed the Neuro-Fuzzy Tornado Detection Algorithm (NFTDA) using a neuro-adaptive system, paving a new path for tornado detection. Hill et al. [15] applied the random forest algorithm to extreme weather warning, and also achieved tornado warning capability. Zeng et al. [16] proposed a tornado detection algorithm based on random forest (TDA-RF), which can effectively identify tornadoes of different intensities. By comprehensively evaluating multiple features, the algorithm demonstrates strong anti-interference capability, enabling tornado warning lead times of up to 18 min based on networked radar. Deep learning has demonstrated strong capability in rapidly identifying various weather patterns but often encounters the problem of data imbalance during the training process. Therefore, Basalyga et al. [17] constructed an image training dataset using tornado data, balanced the training set through image augmentation techniques, and conducted tornado prediction research using an image-based convolutional neural network (CNN) model, further promoting the application of artificial intelligence in the field of tornado warning. A new probabilistic tornado detection algorithm based on azimuthal shear was proposed by Sandmæl et al. [18], which demonstrated superior performance compared to TDA. McGuire [19] successfully predicted the number of tornado outbreak days of different scales based on climate data using three convolutional neural networks: LeNet-5, VGG-16, and ResNet-50. Xie et al. [20] proposed a tornado recognition algorithm based on a multi-task identification network (MTI-Net), introducing a novel multi-head convolutional block backbone network, which effectively improved the tornado recognition hit rate and reduced the false alarm rate, highlighting the effectiveness of MTI-Net in handling small-scale tornado events. A tornado detection algorithm based on XGBoost and dual-polarization radar was proposed by Zeng et al. [21], demonstrating improved performance in tornado identification through the incorporation of dual-polarization parameters.

Recent studies have clearly demonstrated that machine learning and deep learning models have benefited from enhanced computational power, leading to new advancements in tornado detection and early warning, and significantly improving detection and warning efficiency. However, current research in China on tornado structure and evolution remains

https://doi.org/10.3390/rs18081124

largely limited to environmental conditions. Tornado identification models often face a trade-off between increasing detection rates and reducing false alarms due to the lack of high-resolution observational data.

Dual-polarization radar data collected in China, including reflectivity, radial velocity, spectrum width, differential reflectivity, correlation coefficient, and specific differential phase, were utilized to construct a tornado dataset for deep learning applications. A novel tornado detection algorithm, termed TDA-DARKNet (Tornado Detection Algorithm-Dual Attention Residual Dense Kolmogorov–Arnold Network), is proposed. TDA-DARKNet integrates a dual attention mechanism, dense residual connections, and a Kolmogorov–Arnold network (KAN). Specifically, TDA-DARKNet combines the nonlinear function approximation capability of KAN with the spatial feature extraction strengths of convolutional neural networks (CNNs). This design enables more effective representation of nonlinear spatial structures in radar observations. In addition, dense residual connections facilitate efficient integration of shallow detail features and deep semantic features through cross-layer feature reuse and fusion. During feature extraction, a channel-wise attention mechanism (CAM) and a spatial attention mechanism (SAM) are employed to achieve adaptive feature weighting. This process emphasizes critical tornado signatures while suppressing background noise.

2. Dataset

2.1. Dual Polarization Weather Radar

Dual-polarization weather radar extends traditional single-polarization radar by incorporating an additional vertical polarization channel, enabling the measurement of additional parameters, including differential reflectivity (Z_{DR}), correlation coefficient (ρ_{HV}), differential phase (ϕ_{DP}), and specific differential phase (K_{DP}). These parameters can provide more microscopic physical information of the precipitation system, which is of great value in identifying the actual tornado and its debris zone on the ground [22]. The differential reflectivity Z_{DR} refers to the logarithm of the ratio of the reflectivity factors of the horizontal polarization and vertical polarization echoes, reflecting the average shape or flatness of the particles. The definition of Z_{DR} is given in Equation (1). In a normal rainfall environment, the Z_{DR} value of flat raindrops is generally positive, while the non-meteorological debris rolled up by a tornado is usually irregular or nearly spherical, resulting in a significantly lower Z_{DR} value, close to 0 or a negative value.

$$Z_{DR} = 10log(\frac{Z_h}{Z_v}) \tag{1}$$

The correlation coefficient ρ_{HV} represents the consistency of the horizontal and vertical polarization echoes of particles in the radar beam, as shown in Equation (2):

$$\rho_{HV} = \frac{\overline{S_{vv}S_{hh}^*}}{\sqrt{\overline{|S_{vv}|^2}\,\overline{|S_{hh}|^2}}} \tag{2}$$

where S_{hh} and S_{vv} represent the in-phase echoes of horizontally and vertically polarized electromagnetic waves, respectively. When electromagnetic waves continuously illuminate the same resolution volume and the statistical properties of hydrometeors within the volume remain stable, the relative relationship between Z_h and Z_v across pulses remains consistent, resulting in a high correlation coefficient. Correlation coefficients in pure rain regions often exceed 0.98. However, when tumbling, fragmentation, or the coexistence of hydrometeors with differing physical phases (e.g., mixed-phase particles) occurs within the resolution volume, scattering behavior becomes inconsistent. As a result, the relative relationship between Z_h and Z_v across pulses varies significantly, and the correlation

https://doi.org/10.3390/rs18081124

coefficient decreases. Therefore, ρ_{HV} is one of the most critical parameters for identifying the characteristics of tornado debris.

2.2. Radar Base Data Processing

The radar base data used in this study were quality-controlled. During dataset construction, the azimuth indexing was aligned such that the initial radial direction corresponds to geographic North, ensuring consistency between tornado damage survey information (azimuth and distance) and the radar data. This operation is only used for data mapping and does not involve any modification of the original radar measurements.

2.3. Tornado Dataset Construction

There are substantial differences in the intensity, type, and movement paths of tornadoes between China and the United States, with U.S. tornadoes generally exhibiting greater overall intensity. Owing to the relatively late implementation of dual-polarization upgrades to China's new-generation Doppler radar systems, combined with the low frequency of tornado occurrences, the number of available dual-polarization tornado observation samples remains limited. To satisfy the large-sample requirements of deep learning models, while accounting for differences in environmental conditions and radar observation characteristics between Chinese and U.S. tornadoes, radar data from both countries are employed in this study. Specifically, the TorNet dataset [23], developed by Lincoln Laboratory in collaboration with NVIDIA, is integrated with tornado radar observations from China to provide a robust data foundation for model training.

The TorNet dataset facilitates advances in machine learning and deep learning-based tornado identification and detection by providing high-quality simulated radar data. The dataset consists of full-resolution, multi-polarization weather radar data extracted from the National Centers for Environmental Information (NCEI) Storm Events Database spanning 2013–2022, ensuring the reliability and accuracy of the data. In total, the dataset includes 203,133 radar samples, of which confirmed tornado cases account for 6.8% (13,813), false-alarm tornado warnings for 31.8% (64,510), and randomly selected non-tornado weather cases for 61.4% (124,766). This class composition, which maintains a realistic class imbalance and includes challenging non-tornadic cases with tornado-like signatures, enables learning algorithms to better distinguish true tornadoes from non-tornadic convective storms and reduce false positives. The statistical distribution of the dataset is summarized in Figure 1. Collectively, these data provide a comprehensive training resource for deep learning-based tornado detection models.

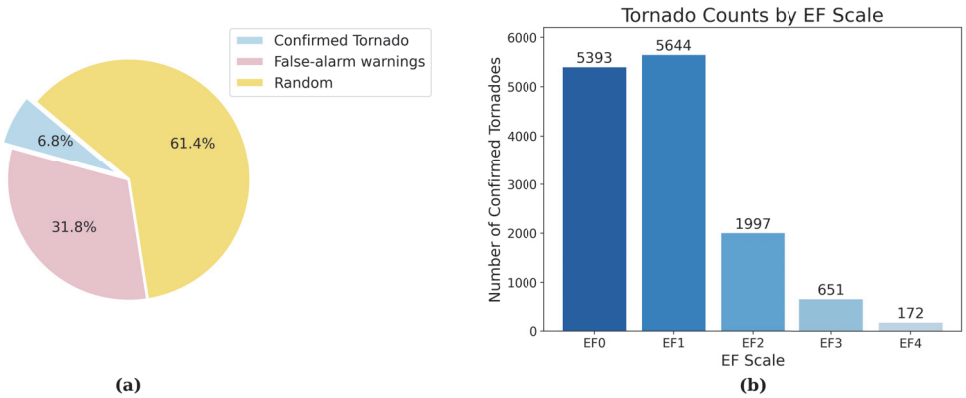


Figure 1. Types and intensities of tornado samples in the Tornet dataset. (**a**) Tornado Sample Counts by Type; (**b**) Distribution of EF-Scale Ratings in Tornado Samples.

Given that tornado cases in China are primarily concentrated in eastern regions, this study selects historical tornado observations from S-band dual-polarization radars

https://doi.org/10.3390/rs18081124

in Jiangsu Province and Guangdong Province. A total of 16 tornado events during the period 2019–2023 were collected. The S-band dual-polarization radar operates in the VCP21 volume coverage pattern, completing a full volume scan every 6 min across multiple elevation angles, including 0.5°, 1.5°, 2.4°, 3.3°, 4.3°, 6.0°, 9.9°, 14.6°, and 19.5°. The radar base data have a radial resolution of 0.25 km, an azimuthal resolution of 1°, and a maximum detection range of 230 km, and are stored in binary format.

The dataset production process is shown in Figure 2. The radar fields were divided into small patches consisting of 8 × 8 range bins (corresponding to a physical area of 2 km × 2 km), due to the small spatial scale of tornadoes and their limited coverage in radar data. A sliding window with a stride of 3 was then applied for sampling to ensure complete coverage of tornado-affected regions. The extracted patches were subsequently filtered and cleaned by removing samples with a high proportion of invalid data, as well as those located near the radar center or at far ranges, where range folding or velocity aliasing effects are significant. Finally, radar variables at the same elevation angle and spatial location, including reflectivity, radial velocity, spectrum width, differential reflectivity, correlation coefficient, and specific differential phase, were reorganized into a data matrix of 6 (data types) × 8 (radials) × 8 (range bins). Positive samples (tornado) were labeled as 1, while negative samples (non-tornado) were labeled as 0.

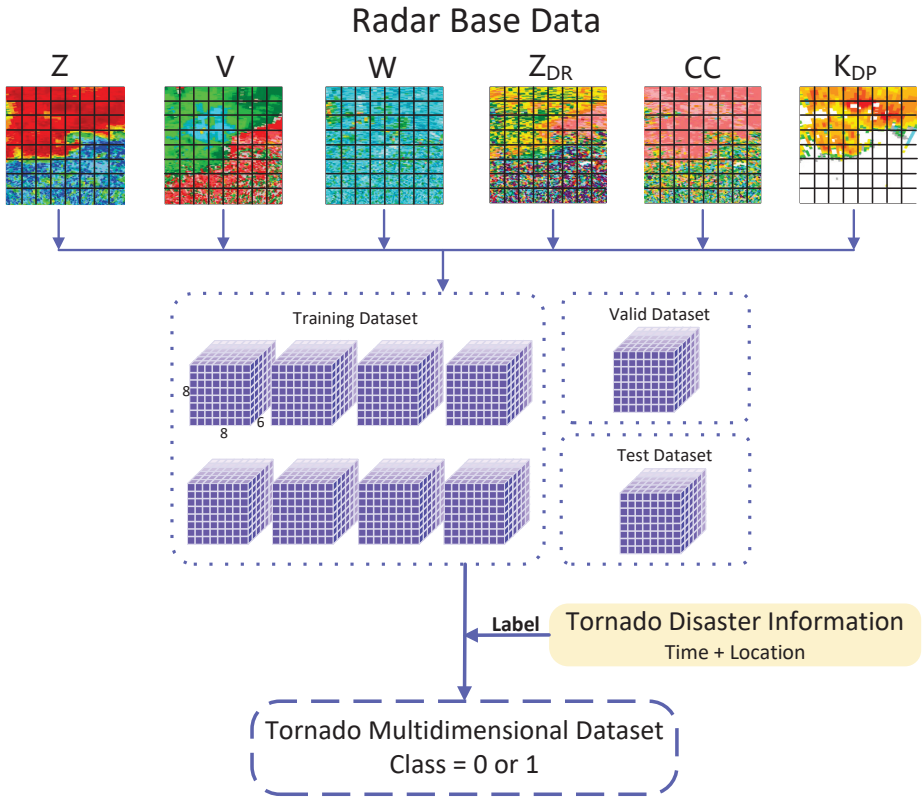


Figure 2. Tornado Sample Set Construction.

A total of 150,177 samples were obtained, consisting of 18,215 positive (tornado) samples and 131,962 negative (non-tornado) samples. To address class imbalance, data augmentation was applied to the positive samples [24]. Specifically, stochastic perturbation-based augmentation methods were employed, including additive Gaussian noise injection, multiplicative amplitude scaling, and small random perturbations. For the negative samples, random undersampling [25] was performed to reduce the number of samples and achieve a more balanced dataset. Consequently, 36,430 positive samples and 36,430 negative samples were retained for model training.

https://doi.org/10.3390/rs18081124

3. Method

3.1. TDA-DARKNet

A novel tornado recognition algorithm, termed Tornado Detection Algorithm-Dual Attention Residual Dense Kolmogorov–Arnold Network (TDA-DARKNet), is proposed in this study. TDA-DARKNet integrates dual attention mechanisms, dense residual structures, and a Kolmogorov–Arnold network (KAN) to improve the accuracy and robustness of tornado identification while reducing the false alarm rate. The attention mechanisms guide the model to focus on critical spatial locations and feature channels associated with tornado occurrence. The dense residual structure enhances feature reuse efficiency, alleviates the vanishing gradient problem, and facilitates deeper extraction of hierarchical tornado echo features. Meanwhile, the KAN component strengthens the model's capability to capture complex nonlinear relationships associated with tornado dynamics.

The architecture of TDA-DARKNet model is illustrated in Figure 3. The input consists of a six-channel radar data tensor, with each channel corresponding to reflectivity, radial velocity, spectrum width, differential reflectivity, correlation coefficient, and specific differential phase. The integrated multi-source tornado dataset, combining the TorNet dataset and the China radar observations, was split into training, validation, and test sets with a ratio of 8:1:1. The training set was used for model optimization, the validation set for monitoring model performance during training and guiding hyperparameter tuning, and the test set for final performance evaluation.

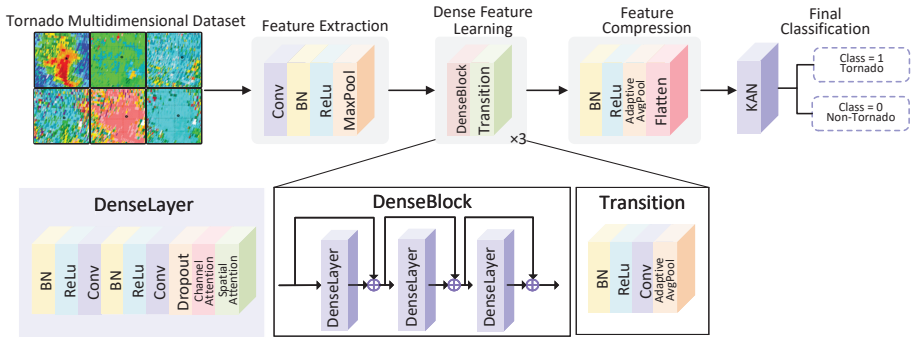


Figure 3. TDA-DARKNet Architecture.

The network backbone is composed of four stages:

Feature Extraction: In this stage, the input radar data patches are first processed by a standard convolutional layer, followed sequentially by batch normalization (BN), a ReLU activation function, and a max-pooling operation, in order to reduce spatial dimensionality and extract preliminary low-level spatial features. This process facilitates the capture of local texture and intensity information associated with tornado occurrence.

Dense Feature Learning: This stage is composed of alternating DenseBlocks and Transition layers. Each DenseBlock consists of multiple DenseLayers and employs dense connectivity, in which each layer receives the outputs of all preceding layers as input, thereby enhancing feature reuse and gradient propagation while improving parameter efficiency [26]. Within each DenseLayer, BN, ReLU activation, Convolution, Dropout, and a dual attention module incorporating both channel and spatial attention are sequentially applied. The dual attention mechanisms embedded within each layer enable adaptive emphasis on critical feature representations closely associated with tornado occurrence, such as rotational signatures or debris-related features. The Transition layers between DenseBlocks perform channel and spatial compression via convolution and adaptive average pooling, which reduce computational cost while preserving salient feature representations,

https://doi.org/10.3390/rs18081124

thereby facilitating the modeling of multiscale tornado features, such as low-level rotational structures and hook echoes.

Feature Compression: After the dense feature learning stage, the feature maps are first sequentially processed by BN and ReLU activation function, and are then compressed to a fixed spatial size through adaptive average pooling to obtain a compact global feature representation. Subsequently, the compressed feature maps are flattened into a one-dimensional vector, which serves as high-level semantic features and is fed into the subsequent stage.

Final Classification: The flattened feature vector is fed into Kolmogorov–Arnold Network (KAN), which replaces the conventional fully connected layer. This module captures the complex nonlinear relationships between radar variables and the probability of tornado occurrence through spline-based basis function representations. Finally, the model outputs a binary classification probability indicating whether a tornado is present in the given sample.

3.2. Kolmogorov–Arnold Network

Kolmogorov–Arnold Network (KAN) is a neural network architecture derived from the Kolmogorov–Arnold representation theorem [27]. The Kolmogorov–Arnold representation theorem was proposed by Vladimir Arnold and Andrey Kolmogorov. It states that a multivariate continuous function defined on a bounded domain can be represented as a composition of finitely many univariate continuous functions and binary addition operations, as shown in Equation (3):

$$f(X) = f(x_1, \ldots, x_n) = \sum_{q=1}^{2n+1} \Phi_{\mathrm{q}} \left(\sum_{\mathrm{p}=1}^{n} \Phi_{\mathrm{q,p}}(x_p) \right) \tag{3}$$

where $\Phi_q : \mathbb{R} \to \mathbb{R}, \Phi_{q,p} : [0,1] \to \mathbb{R}, x_p$ represents the pth element of vector X.

In KAN, network weights are not represented as single scalar values. Instead, they are modeled using B-spline basis functions, and weighted summation is performed only at spline knots. Multilayer perceptrons (MLPs) define weights as real-valued parameters and rely on fixed activation functions to introduce nonlinearity [28]. In contrast, KAN can directly learn complex functional relationships. It has demonstrated high accuracy and good interpretability in small-sample scientific modeling tasks. KAN is structurally similar to MLPs at the network level. Internally, it is consistent with spline-based representations. This design enables KAN to effectively learn tornado-related features and to model the extracted features with high precision.

Within the KAN framework, B-spline functions are employed to implement the univariate mapping module. This design preserves the expressive capacity of the network and improves its stability and interpretability in small-sample modeling. Compared with global polynomial fitting, B-splines provide clear advantages in modeling nonlinear relationships and capturing local features [29,30]. The construction of B-spline functions begins with the zeroth-order ($k = 1$) basis function. This function takes a value of 1 within the interval defined by two adjacent knots and 0 outside the interval, thereby exhibiting strict local support, as shown in Equation (4). Based on this, higher-order ($k \geq 2$) B-spline basis functions are recursively derived using the de Boor-Cox formula. Each higher-order basis function is expressed as a linear combination of two adjacent lower-order basis functions (Equation (5)). This formulation achieves higher-order continuity and smoothness while preserving locality. Finally, a continuous and smooth function with local control properties

https://doi.org/10.3390/rs18081124

is constructed through the weighted summation of control coefficients and basis functions, as shown in Equation (6).

$$B_{i,1}(x) = \begin{cases} 1, & t_i \le x \le t_{i+1} \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

$$B_{i,k}(x) = \frac{x - t_i}{t_{i+k-1} - t_i} B_{i,k-1}(x) + \frac{t_{i+k} - x}{t_{i+k} - t_{i+1}} B_{i+1,k-1}(x) \tag{5}$$

$$S(x) = \sum_{i=0}^{n} c_i B_{i,k}(x) \tag{6}$$

3.3. Channel Attention Mechanism

The Channel Attention Mechanism (CAM) is introduced to enable the model to automatically emphasize more informative feature subspaces. CAM [31] is designed to allocate importance across feature map channels by assigning adaptive weights to each tornado-related channel. This mechanism highlights features that contribute most to recognition while suppressing irrelevant or redundant information, thereby enhancing overall model performance.

As shown in Figure 4, the channel attention mechanism typically consists of the following three steps (taking the input feature $X \in \mathbb{R}^{C \times H \times W}$ as an example):

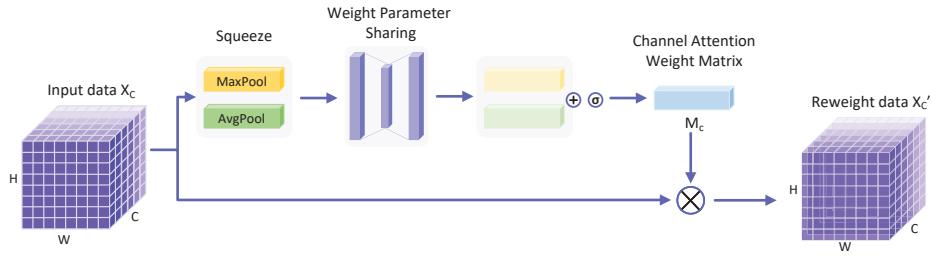


Figure 4. Channel Attention Mechanism.

(1) Channel Feature Compression: Global average pooling (GAP) and global max pooling (GMP) are first applied to compress the spatial dimensions, converting the spatial information of each channel into a single scalar. This process yields a channel-wise descriptor vector, as shown in Equation (7).

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_c(i, j) \tag{7}$$

(2) Channel Attention Weight Calculation: The channel descriptor vector is fed into a lightweight fully connected neural network , typically a two-layer perceptron, to capture inter-channel dependencies and compute the attention weights for each channel, as shown in Equation (8).

$$s = \sigma(W_2 \cdot \mathrm{ReLU}(W_1 \cdot z)) \tag{8}$$

Here, σ denotes the Sigmoid function, and the output $s \in [0, 1]^C$ represents the weights (attention coefficients) of each channel.

(3) Feature Recalibration: Finally, each channel in the original feature map is multiplied by its corresponding weight through channel-wise multiplication, thereby completing the weighted recalibration of channels.

$$X_c' = s_c \cdot X_c \tag{9}$$

https://doi.org/10.3390/rs18081124

3.4. Spatial Attention Mechanism

The Spatial Attention Mechanism (SAM) is an attention module in neural networks that dynamically assigns importance to spatial positions [30]. SAM aims to guide the model in automatically identifying spatial regions that require greater "attention", thereby emphasizing key areas associated with tornado occurrence and suppressing irrelevant background, ultimately enhancing the model's feature representation capability.

As shown in Figure 5, the spatial attention mechanism typically consists of the following three steps (taking the input feature $X \in \mathbb{R}^{C \times H \times W}$ as an example):

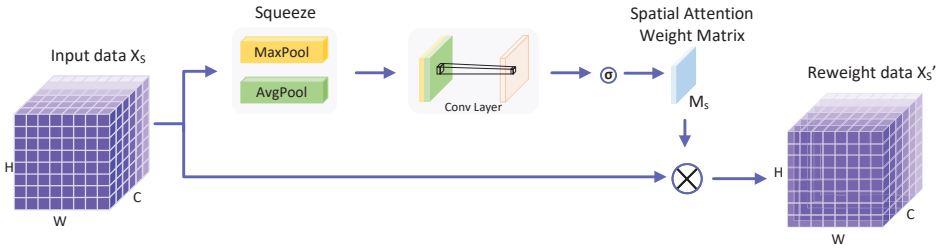


Figure 5. Spatial Attention Mechanism.

(1) Channel Dimension Compression: Global max pooling and global average pooling are first applied along the channel dimension of an input feature X with size C × H × W, producing two feature maps of size 1 × H × W, as shown in Equations (10) and (11).

$$X_{\max} = \mathrm{MaxPool}_c(X) \tag{10}$$

$$X_{\mathrm{avg}} = \mathrm{AvgPool}_c(X) \tag{11}$$

(2) Spatial Attention Weight Calculation: Subsequently, the two pooled feature maps are concatenated along the channel dimension and passed through a convolutional layer with a kernel size of 7 × 7. Finally, a Sigmoid activation function is applied to generate the spatial attention weight map M_s.

$$M_s = \sigma\left(f^{7\times7}\left(\left[X_{\max}; X_{\mathrm{avg}}\right]\right)\right) \tag{12}$$

Here, $f^{7\times7}(\cdot)$ denotes the convolution operation and represents the Sigmoid function.

(3) Feature Recalibration: Finally, each spatial element in the original feature map is multiplied element-wise by its corresponding weight, completing the spatial-wise weighted processing, as shown in Equation (13).

$$X' = M_s \otimes X \tag{13}$$

3.5. Model Evaluation

When evaluating binary classification models in deep learning algorithms, a binary confusion matrix is commonly used to assess model performance, as shown in Table 1. TP denotes the number of correctly classified positive samples, corresponding to correctly identified tornadoes. FP denotes the number of incorrectly classified positive samples, corresponding to falsely identified tornadoes. FN denotes the number of incorrectly classified negative samples, corresponding to missed tornadoes. TN denotes the number of correctly classified negative samples, corresponding to correctly identified non-tornado cases.

https://doi.org/10.3390/rs18081124

Table 1. Confusion Matrix for Binary Classification.

Ground Truth	Predict	
	Positive (Tornado)	**Negative (Non-Tornado)**
Positive (Tornado)	*TP*	*FN*
Negative (Non-Tornado)	*FP*	*TN*

Based on the confusion matrix, several evaluation metrics are used to assess model performance, including accuracy (*ACC*; Equation (14)), precision (*PRE*; Equation (15)), F1-score (Equation (17)), probability of detection (*POD*; Equation (18)), false alarm ratio (*FAR*; Equation (19)), critical success index (*CSI*; Equation (20)), and Matthews correlation coefficient (*MCC*; Equation (21)).

$$ACC = \frac{TP + TN}{TP + FN + FP + TN} \tag{14}$$

$$PRE = \frac{TP}{TP + FP} \tag{15}$$

$$Recall = \frac{TP}{TP + FN} \tag{16}$$

$$F_{1-\text{ score}} = \frac{2 \cdot PRE \cdot Recall}{PRE + Recall} \tag{17}$$

$$POD = \frac{TP}{TP + FN} \tag{18}$$

$$FAR = \frac{FP}{TP + FP} \tag{19}$$

$$CSI = \frac{TP}{TP + FN + FP} \tag{20}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + TP)(TN + FN)}} \tag{21}$$

In order to comprehensively evaluate the performance of the model, the experiment selected five common deep learning classification models to compare with the model TDA-DARKNet proposed in this paper, including ResNet, DenseNet, EfficientNet, ConvNeXt, and MLP-Mixer model. The evaluation results on the test set are summarized in Table 2.

Table 2. Evaluation metric scores of the model on the test set, where "↑" indicates that a higher score is better, and "↓" indicates that a lower score is better.

Model	ACC ↑	PRE ↑	F1-Score ↑	POD ↑	FAR ↓	CSI ↑	MCC ↑
ResNet	0.9550	0.9367	0.9559	0.9759	0.0633	0.9155	0.9108
DenseNet	0.9647	0.9419	0.9655	0.9904	0.0581	0.9333	0.9306
EfficientNet	0.9651	0.9413	0.9606	0.9807	0.0586	0.9242	0.9204
ConvNeXt	0.9606	0.9449	0.9686	0.9935	0.0550	0.9392	0.9369
MLP-Mixer	0.9654	0.9227	0.9501	0.9790	0.0772	0.9049	0.8988
TDA-DARKNet	0.9711	0.9508	0.9717	0.9936	0.0492	0.9450	0.9431

The experimental results indicate that TDA-DARKNet outperforms all comparative models in the tornado identification task. The model achieves an ACC of 0.9711, with PRE and F1-score reaching 0.9508 and 0.9717, respectively, demonstrating high accuracy and

https://doi.org/10.3390/rs18081124

strong performance balance. Its POD reaches 0.9936, indicating a significantly reduced miss rate, while the FAR is as low as 0.0492, reflecting strong false alarm suppression capability. In addition, the CSI and MCC attain 0.9450 and 0.9431, respectively, further verifying the model's stability and discriminative ability under complex radar interference conditions. Overall, TDA-DARKNet exhibits superior performance across key classification and warning metrics, demonstrating strong potential for operational application.

To evaluate the influence of each module on model performance, ResDenseNet was used as the baseline model, and different variants were constructed by sequentially introducing channel attention module (CA), spatial attention module (SA), convolutional block attention module (CBAM), and the KAN module. The effects of individual modules were first compared, followed by an analysis of the performance changes when attention mechanisms were combined with the KAN module. All experiments were conducted under the same data split and training settings to ensure fairness. This design enables a quantitative assessment of the contribution of each module and verifies the effectiveness of the final TDA-DARKNet model.

The ablation study results (Table 3) systematically evaluate the contributions of different modules to the model performance. The baseline demonstrates strong detection capability (POD = 0.9904). However, it suffers from a relatively high false alarm rate (FAR = 0.0667), indicating limitations in distinguishing strongly rotating but non-tornadic cases. The introduction of CA or SA does not lead to consistent performance improvements, and even results in degradation in some metrics. For example, incorporating CA increases the FAR to 0.0790, suggesting that a single attention mechanism may introduce redundant features or amplify non-essential echo information. Similarly, the CBAM module (CA + SA) fails to provide noticeable gains, indicating that conventional attention mechanisms have limited effectiveness in tornado identification tasks. In contrast, incorporating the KAN module leads to substantial performance improvements (ACC increases from 0.9598 to 0.9679, while FAR decreases to 0.0495), demonstrating its superiority in modeling complex nonlinear relationships. The proposed TDA-DARKNet achieves the best performance across all evaluation metrics, while maintaining an extremely high detection rate (POD = 0.9936) and reducing the false alarm rate (FAR = 0.0492), thereby validating the rationality and effectiveness of the proposed model architecture.

Table 3. Ablation study results of different model configurations, where "↑" indicates that a higher score is better, and "↓" indicates that a lower score is better.

Model	ACC ↑	PRE ↑	F1-Score ↑	POD ↑	FAR ↓	CSI ↑	MCC ↑
ResDenseNet	0.9598	0.9333	0.9610	0.9904	0.0667	0.9249	0.9214
ResDenseNet + CA	0.9542	0.9210	0.9559	0.9936	0.0790	0.9156	0.9113
ResDenseNet + SA	0.9582	0.9425	0.9589	0.9759	0.0575	0.9211	0.9170
ResDenseNet + CBAM	0.9574	0.9344	0.9585	0.9839	0.0656	0.9203	0.9162
ResDenseNet + KAN	0.9679	0.9505	0.9685	0.9871	0.0495	0.9388	0.9364
ResDenseNet + CA + KAN	0.9614	0.9375	0.9624	0.9887	0.0625	0.9276	0.9243
ResDenseNet + SA + KAN	0.9622	0.9376	0.9633	0.9904	0.0624	0.9291	0.9260
TDA-DARKNet	0.9711	0.9508	0.9717	0.9936	0.0492	0.9450	0.9430

The channel and spatial attention module can adaptively assign importance weights to different features. Specifically, channel attention emphasizes the relative contributions of key meteorological variables (e.g., velocity and reflectivity variables), while spatial attention focuses on critical regions within severe convective systems that exhibit pronounced dynamical or microphysical characteristics (e.g., high reflectivity regions). The Kolmogorov–Arnold Network (KAN) explicitly learns nonlinear functional relationships

https://doi.org/10.3390/rs18081124

between input features and outputs. Compared with traditional black-box neural networks, its functional form is more interpretable. This structure enables the model to more clearly characterize the nonlinear response relationships between the key features highlighted by the channel and spatial attention module and tornado occurrence, thereby revealing potential underlying physical connections to some extent.

Considering that performance differences among deep learning models on this dataset are relatively limited, comparisons restricted solely to deep learning architectures may not fully reflect the effectiveness of algorithmic improvements. Therefore, a representative machine learning model, TDA-XGB [21], is further introduced as a baseline to evaluate performance enhancement from a different technical perspective, as shown in Table 4. TDA-XGB is a tornado identification approach based on Extreme Gradient Boosting (XGBoost). Its core principle lies in constructing a high-accuracy strong classifier through an ensemble learning framework that sequentially aggregates multiple weak learners (CART regression trees). The method minimizes an objective function composed of a loss term and a regularization term, thereby improving predictive accuracy while effectively controlling model complexity and reducing the risk of overfitting. The model is trained iteratively using a gradient boosting strategy, where each round takes the residuals of the current model as the learning target to progressively refine the decision boundary and ultimately output the probability of tornado occurrence. In terms of feature construction, TDA-XGB relies on manual feature engineering based on dual-polarization radar data. It extracts statistical descriptors, including the maximum, minimum, and mean values, of reflectivity, radial velocity, spectrum width, differential reflectivity, and correlation coefficient. In addition, dynamic features such as velocity difference, velocity shear, angular momentum, and rotational velocity are derived to construct a multidimensional feature vector for classification.

Table 4. Evaluation metric scores of TDA-XGB and TDA-DARKNet on the test set, where "↑" indicates that a higher score is better and "↓" indicates that a lower score is better.

Model	ACC ↑	PRE ↑	F1-Score ↑	POD ↑	FAR ↓	CSI ↑	MCC ↑
TDA-XGB	0.8387	0.8222	0.8402	0.8590	0.1778	0.7244	0.6782
TDA-DARKNet	0.9711	0.9508	0.9717	0.9936	0.0492	0.9450	0.9431

Overall, machine learning approaches typically rely on manually engineered statistical features or predefined physical criteria. As a result, their capability to characterize the complex spatial structures and weak rotational signatures associated with tornadoes is relatively limited, making them more prone to missed detections in weak tornado cases or during the early developmental stages. In contrast, the proposed TDA-DARKNet demonstrates higher detection sensitivity in the tornado identification task. This improvement can be attributed to its deep feature learning framework, which is capable of extracting more discriminative spatial structural information directly from dual-polarization radar base data and enhancing the response to critical tornado echo regions. Consequently, the model maintains strong identification performance even during the early development phase of tornadoes and under weak-signal conditions. Furthermore, with the incorporation of the Kolmogorov–Arnold Network (KAN), the model achieves a more comprehensive approximation of complex nonlinear relationships, thereby further improving overall recognition robustness.

4. Typical Case Analysis and Generalization Capability Validation

Due to the pronounced case-to-case variability and sudden nature of tornadoes, a reliable model must not only achieve strong performance on evaluation metrics but also

https://doi.org/10.3390/rs18081124

maintain stable identification capability across typical events and varying observational conditions. Therefore, this section conducts validation analyses using multiple independent tornado cases from Jiangsu and Guangdong Provinces. Specifically, two representative tornado events are selected to evaluate the ability of TDA-DARKNet to identify tornadoes of different intensities, including an EF2 tornado that occurred in Guangzhou, Guangdong, on 27 April 2024, and an EF0 tornado on 17 September 2024. On this basis, statistical identification results from 19 tornado events are further analyzed to comprehensively assess the overall generalization capability of the proposed model. The tornado damage survey data used in this study are obtained from local meteorological agencies and are derived from field investigations combined with multiple data sources, which are used to determine the occurrence time, path, and touchdown location of tornadoes.

4.1. EF2 Tornado in Guangzhou, Guangdong on 27 April 2024

At 15:00 on 27 April 2024, a tornado struck Liangtian Village in Zhongluotan Town, Baiyun District, Guangzhou, Guangdong Province. According to the damage survey, this event had a path length of approximately 8.2 km and an average width of about 700 m. The tornado was classified as an EF2 event, indicating a strong tornado. Eight volume scans from the Z9200 radar between 14:18 and 15:00 were used to evaluate this tornado case. The model identification results are marked by black rectangles on the radar Plan Position Indicator (PPI) images.

The TDA-DARKNet model first detected early tornado signatures at 14:18. As shown in Figure 6, hook echo signatures exhibited a clear evolutionary pattern across all elevation angles in the reflectivity PPI images.

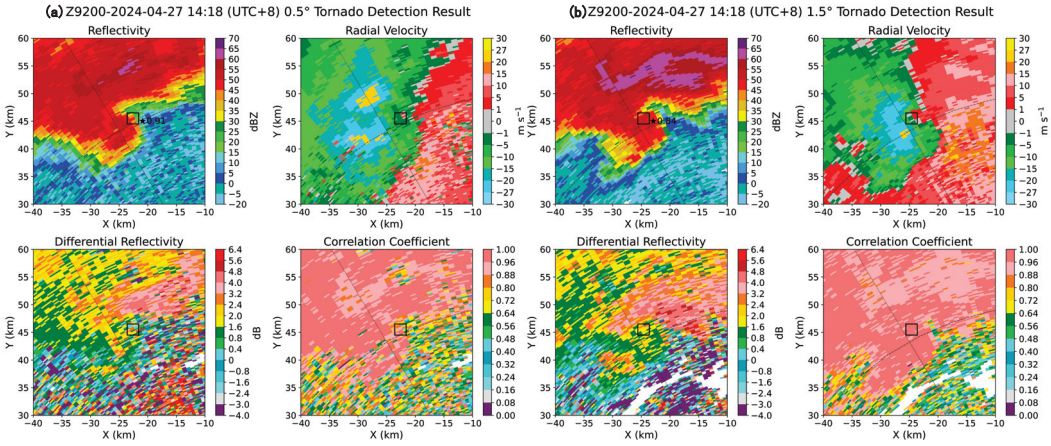


Figure 6. Detection results of TDA-DARKNet from the Z9200 radar at 14:18 (UTC + 8). The black rectangle is centered on the detected location and represents a square region with a side length of 2 km, while the annotated values indicate the corresponding detection probabilities. Sub-images (**a**,**b**) correspond to elevation angles of 0.5° and 1.5°, respectively. Each sub-image consists of four radar variables, including reflectivity, radial velocity, differential reflectivity, and correlation coefficient, arranged in a 2 × 2 layout.

Figure 7 illustrates the identification results of TDA-DARKNet at 15:00. According to damage survey records, the tornado had already touched down by this time. The identified region exhibits high reflectivity, accompanied by pronounced positive-negative radial velocity couplets. Low values of differential reflectivity and correlation coefficient are also observed. These features indicate a typical TDS in the radar PPI images.

https://doi.org/10.3390/rs18081124

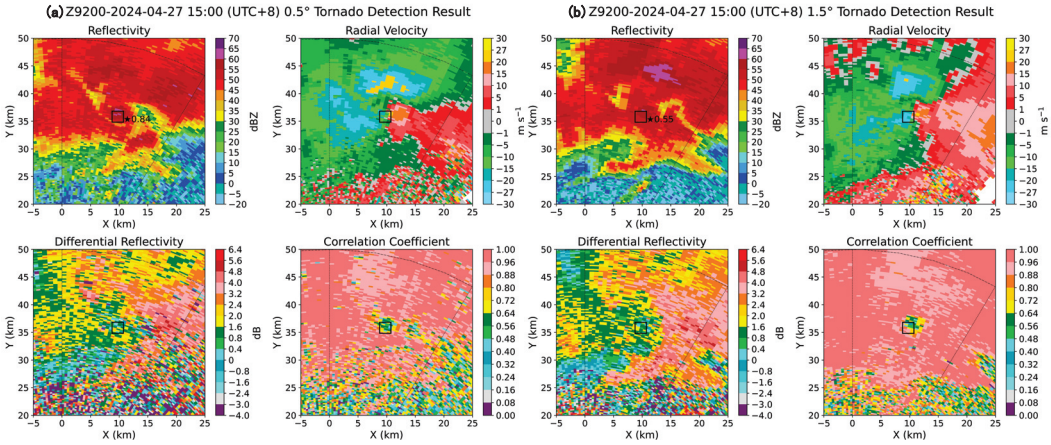


Figure 7. Detection results of TDA-DARKNet from the Z9200 radar at 15:00 (UTC + 8). The black rectangle is centered on the detected location and represents a square region with a side length of 2 km, while the annotated values indicate the corresponding detection probabilities. Sub-images (**a**,**b**) correspond to elevation angles of 0.5° and 1.5°, respectively. Each sub-image consists of four radar variables, including reflectivity, radial velocity, differential reflectivity, and correlation coefficient, arranged in a 2 × 2 layout.

In this case, the TDA-XGB model first produced a valid detection at 14:48, providing a lead time of approximately 12 min. In contrast, TDA-DARKNet was able to capture potential tornado development features at an earlier stage, achieving a lead time of up to 42 min. The feature-engineered TDA-XGB relies on regional statistical metrics or threshold-based criteria as input variables. During feature construction, part of the spatial structural information contained in radar echoes is inevitably compressed. As a result, its capability to characterize fine-scale morphological features, such as mesocyclones and hook echoes, is relatively limited, making it difficult to identify tornado signatures during the early formation stage or when the signals are not yet fully developed.

As shown in Figure 8, we analyzed the tornado positions identified by TDA-DARKNet at the same elevation between 14:18 and 15:00. The tornado movement path identified by the model exhibits an overall west to east-southeast trajectory, which is generally consistent with the post-event damage survey results. This agreement indicates that TDA-DARKNet is capable of accurately capturing the spatiotemporal evolution of the tornado.

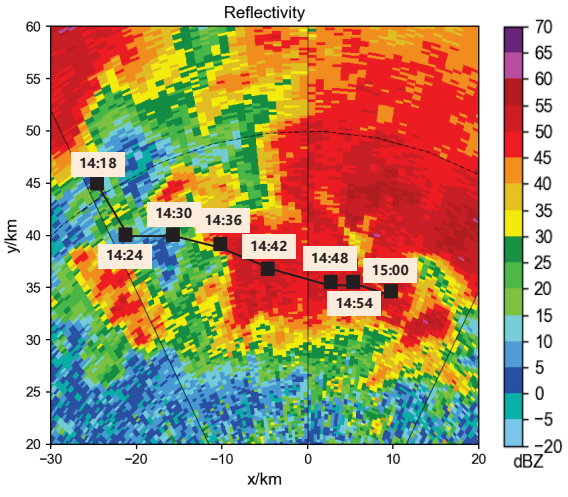


Figure 8. Detection-based tornado path of the EF2 tornado in Guangzhou on 27 April 2024 using TDA-DARKNet.

https://doi.org/10.3390/rs18081124

4.2. EF0 Tornado in Xuzhou, Jiangsu on 17 September 2024

At 12:21, the TDA-DARKNet model detected tornado signatures at the 1.5° and 2.4° elevation angles, as shown in Figure 9. At the 1.5° elevation, the detected region exhibited high reflectivity. The positive-negative radial velocity couplets reached 27 $m \cdot s^{-1}$ and -22.5 $m \cdot s^{-1}$, respectively. These features indicate a typical tornado vortex signature.

Figure 10 shows the identification results of the Z9516 radar at 12:32 (UTC + 8) for the 1.5° and 2.4° elevations. This scan can be regarded as the radar observation closest to the touchdown time. As shown in the figure, the detected regions at the 1.5° elevation exhibit high reflectivity. Hook echo signatures and the corresponding positive-negative velocity couplets are clearly visible.

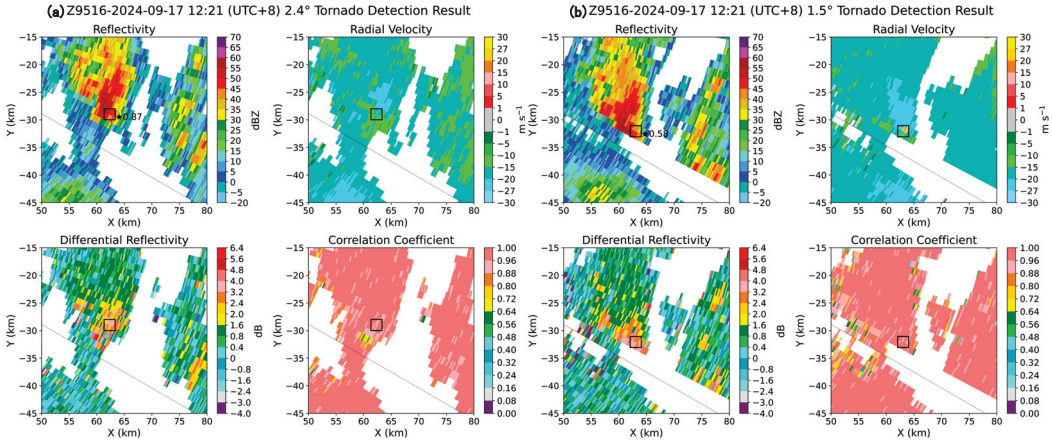


Figure 9. Detection results of TDA-DARKNet from the Z9516 radar at 12:21 (UTC + 8). The black rectangle is centered on the detected location and represents a square region with a side length of 2 km, while the annotated values indicate the corresponding detection probabilities. Sub-images (**a**,**b**) correspond to elevation angles of 1.5° and 2.4°, respectively. Each sub-image consists of four radar variables, including reflectivity, radial velocity, differential reflectivity, and correlation coefficient, arranged in a 2 × 2 layout.

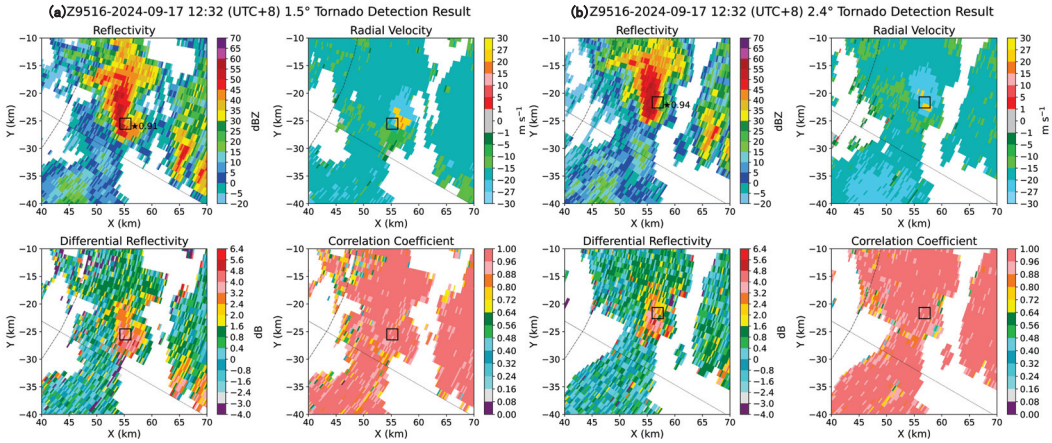


Figure 10. Detection results of TDA-DARKNet from the Z9516 radar at 12:32 (UTC + 8). The black rectangle is centered on the detected location and represents a square region with a side length of 2 km, while the annotated values indicate the corresponding detection probabilities. Sub-images (**a**,**b**) correspond to elevation angles of 1.5° and 2.4°, respectively. Each sub-image consists of four radar variables, including reflectivity, radial velocity, differential reflectivity, and correlation coefficient, arranged in a 2 × 2 layout.

https://doi.org/10.3390/rs18081124

In this weak tornado case, TDA-XGB failed to provide a valid detection result, whereas TDA-DARKNet successfully identified tornado signatures 9 min in advance. This outcome indicates that the deep learning approach, by enhancing the representation of small-scale spatial structures in radar echoes and capturing complex nonlinear relationships, exhibits greater sensitivity to weak rotational features and early-stage tornado signals. Consequently, it demonstrates clear advantages in terms of tornado detection probability and warning lead time.

4.3. Generalization Capability Validation

To further evaluate the generalization capability of the proposed model under different regional and radar observational conditions, a total of 19 independent tornado events from Jiangsu, Guangdong, and Anhui Provinces were selected for validation, as summarized in Table 5. In the table, the Detection value of 1 indicates that a tornado event was successfully identified by the model, while the False Alarm value of 1 indicates that the model issued a false alarm. When both Detection and False Alarm values are equal to 1, it indicates that the model issued multiple tornado warnings, including at least one correct detection (hit) along with additional false alarms. Considering that model performance is sensitive to input data quality, rigorous quality control and screening procedures were conducted prior to the generalization assessment. Cases with severe missing data, significant ground clutter contamination, or substantial influence from radar beam blockage and detection blind zones were excluded to ensure the reliability and comparability of the statistical results. It should be noted that for cases with relatively poor data quality, the model's detection performance may still be affected to some extent. This observation further highlights the critical importance of high-quality observational data in deep learning-based tornado identification tasks.

Table 5. Statistics of detection results and warning lead times for tornado events in different regions. Detection = 1 indicates successful detection; False Alarm = 1 indicates false alarm; "/" indicates no record.

Date	Time (UTC + 8)	Location	EF Scale	Radar	Detection	False Alarm	Lead Time
20200518	14:06–14:07	Guangdong Jiangmen	EF0	Z9662	0	1	/
20200601	12:50–12:57	Guangdong Guangzhou	/	Z9200	1	0	32 min
20200612	13:49–14:00	Jiangsu Gaoyou	EF2	Z9250	1	0	13 min
20200722	21:48	Jiangsu Suqian	EF3	Z9515	1	1	18 min
20200722	22:40	Jiangsu Yancheng	EF2	Z9515	1	0	25 min
20210513	14:40	Guangdong Zhuhai	/	Z9755	0	0	/
20210728	16:30–18:40	Guangdong Guangzhou	/	Z9200	1	0	6 min
20210728	16:30–18:40	Guangdong Guangzhou	/	Z9758	1	0	6 min
20210730	10:00–11:00	Guangdong Guangzhou	/	Z9200	1	0	6 min
20220619	07:20–07:30	Guangdong Foshan	EF1	Z9758	1	0	10 min
20230610	16:22	Jiangsu Nantong	EF2	Z9513	1	0	15 min
20240427	14:49–15:04	Guangdong Guangzhou	EF2	Z9200	1	0	42 min
20240917	12:30–12:38	Jiangsu Xuzhou	EF0	Z9516	1	0	9 min
20240917	14:07–14:12	Jiangsu Xuzhou	EF1	Z9516	1	1	7 min
20240917	14:33–14:41	Jiangsu Xuzhou	EF1	Z9516	1	0	11 min
20240917	15:31–15:33	Jiangsu Xuzhou	EF0	Z9516	1	0	2 min
20240917	15:47–16:10	Jiangsu Lianyungang	EF0	Z9516	1	0	7 min
20240917	16:02–16:09	Jiangsu Xuzhou	EF0	Z9516	1	0	5 min
20240917	17:07–17:16	Anhui Suzhou	EF1	Z9516	1	0	5 min

The statistical results indicate that TDA-DARKNet provides effective detections in the majority of cases, maintaining a relatively high overall detection rate. The number of

https://doi.org/10.3390/rs18081124

false alarm cases is limited, suggesting that the model retains good stability and robustness under complex severe convective conditions. In terms of warning lead time, the model achieves varying degrees of advance detection in all successfully identified cases, with an average lead time of approximately 13 min, demonstrating strong potential for operational application. For most events, the lead time is concentrated within the 5–15 min range, while in several intense tornado cases, the advance warning time exceeds 40 min. This result indicates that the model is capable of capturing key structural and dynamical features during the early stages of tornado development.

Regarding intensity distribution, the generalization test samples cover tornadoes ranging from EF0 to EF3. The model not only consistently identifies strong tornado events of EF2 and above, but also exhibits reliable detection capability for weaker EF0–EF1 tornado cases. This finding suggests that the deep learning approach maintains strong feature extraction and discrimination ability even under weak rotational signal conditions. Cross-regional statistical analysis further demonstrates that the model achieves generally consistent detection performance under different radar observational conditions in Jiangsu and Guangdong Provinces, highlighting its regional adaptability and generalization capability.

Overall, the multi-event statistical results confirm that TDA-DARKNet maintains strong detection stability and generalization performance across different intensity levels, regional environments, and radar observational conditions. However, this conclusion still requires further validation across a broader range of regions and more diverse sample conditions. Furthermore, the implementation of rigorous data quality control procedures provides a solid foundation for reliable model performance evaluation and underscores the practical necessity of high-quality radar data support for future operational deployment.

5. Discussion

In the EF0 tornado case that occurred in Xuzhou in September 2024, the TDA-DARKNet model first detected tornado features at 12:21, providing approximately 9 min of lead time. A comparison of the two tornado cases indicates that TDA-DARKNet provides substantially earlier warnings for the EF2 strong tornado than for the EF0 weak tornado.

This difference is primarily attributed to more pronounced and longer-lasting mesocyclone echo signatures during the early stages of strong tornadoes. These signatures allow the TDA-DARKNet model to capture key features earlier, thereby improving detection accuracy and response timeliness. EF0 tornadoes are generally smaller in spatial scale and shorter in duration. Their structural features are often less distinct and less stable than those of strong tornado systems. As a result, weak tornadoes often lack clear mesocyclone signatures or classic echo structures in radar base data. This limitation increases detection difficulty and reduces early-warning effectiveness.

In the Guangzhou event, tornado debris signature (TDS) features became clearly identifiable approximately two to three volume scans before touchdown. However, in the Xuzhou EF0 case, TDS features were rarely observed throughout the entire tornado life cycle. This result suggests that weak tornadoes typically generate limited near-surface debris, making TDS-based detection challenging. This indicates that TDS has limited utility for advance tornado warnings, especially for weak tornado events. However, it remains a highly reliable indicator for confirming tornado touchdown. Therefore, early warning of weak tornadoes should rely more heavily on other radar features, particularly strong low-level rotational signatures, such as tornado vortex signatures (TVS) or near-surface mesocyclones.

From the preceding discussion, the following findings can be drawn:

1. Radar remains the primary tool for tornado detection, but it has inherent limitations. Radar can identify potential or ongoing tornadoes by detecting mesocyclones and tornado

https://doi.org/10.3390/rs18081124

vortex signatures (TVS) within thunderstorms. However, not all detected vortices result in tornado touchdown. The detection of a mesocyclone does not necessarily indicate that a tornado will reach the ground. Issuing warnings for every mesocyclone would lead to excessive false alarms, potentially triggering a "cry wolf" effect among the public. Moreover, not all tornadoes exhibit clear radar signatures. Some tornadoes, particularly non-supercell tornadoes, have very short lifespans and weak or indistinct rotational signatures. In addition, radar observations are constrained by temporal and spatial resolution, as scans are performed at discrete time intervals. A tornado may form and cause damage between successive scans. The integration of high-resolution phased-array radar systems with environmental field nowcasting represents a key direction for improving tornado detection and early warning.

2. The average lifespan of a tornado is approximately 10–20 min. The process from radar identification to warning issuance must be completed within a very limited time window. In the United States, the average effective tornado warning lead time is approximately 13–15 min. In China, effective warning lead times are often shorter, due to the smaller scale and rapid evolution of tornadoes. Tornado warnings must clearly specify the threatened areas. However, the highly unpredictable nature of tornado paths complicates accurate impact-area forecasting. As a result, warning regions may be overly broad or insufficiently targeted. Effective tornado warning requires the integration of detection and path prediction algorithms, together with rapid information dissemination through mobile emergency alerts and social media platforms, to ensure timely public response.

6. Conclusions

An improved deep learning model, TDA-DARKNet, was proposed for tornado identification based on radar base data. The model integrates dual attention mechanisms, dense residual structures, and the Kolmogorov–Arnold Network (KAN), and demonstrates strong performance in improving detection accuracy, reducing false alarm rates, and enhancing the recognition capability for weak tornadoes. The main findings are summarized as follows:

1. TDA-DARKNet integrates the spatial feature extraction capabilities of convolutional neural networks with both channel and spatial attention mechanisms to strengthen the response to key variables and spatial regions. Additionally, the KAN module enhances its capacity to model complex nonlinear relationships. TDA-DARKNet achieved an accuracy of 97.11% and a precision of 99.36% on the test set, with a false alarm ratio of only 4.92%, which effectively mitigates the trade-off between accuracy and false alarm rate in tornado detection.

2. The typical case analyses further demonstrate the practical application potential of the proposed model across tornadoes of different intensities. In the strong tornado event on 27 April 2024, TDA-DARKNet achieved a maximum warning lead time of 42 min, and the identified movement path was generally consistent with the post-event damage survey results. In the weak tornado case, the model successfully provided a valid detection approximately 9 min in advance, whereas TDA-XGB failed to identify the event and resulted in a missed detection. These results indicate that the deep learning approach exhibits higher detection sensitivity under weak rotational or non-typical structural conditions.

3. The cross-regional multi-event generalization results further indicate that, across 19 independent tornado cases from Jiangsu, Guangdong, and Anhui Provinces, the model maintains a consistently high overall detection rate and demonstrates stable performance under varying radar observational conditions. Most successfully detected events achieved minute-level advance warnings, with lead times exceeding 40 min in several strong tornado cases. Although a small number of false alarms were observed, the overall false alarm level remains effectively controlled compared with the relatively high false alarm

https://doi.org/10.3390/rs18081124

rates commonly encountered in current operational tornado detection systems. These findings suggest that TDA-DARKNet exhibits strong detection stability and a certain degree of regional generalization capability across different regions, intensity levels, and observational conditions.

4. The model performs automatic detection directly based on radar base data, eliminating the need for manual feature engineering and thereby simplifying the data processing workflow. Meanwhile, the incorporation of the KAN module enhances the modeling of complex nonlinear relationships, improving the model's generalization performance and robustness under limited sample conditions. The results suggest that the transition from "handcrafted feature-driven" approaches to "deep representation-driven" frameworks constitutes a key technological pathway for improving tornado detection probability and extending warning lead time.

Current tornado identification methods primarily rely on radar echo data. However, dependence on a single data source remains a key limitation. Tornadoes are highly variable and complex, and the limited spatial coverage and viewing geometry of individual radars make it difficult to fully capture their three-dimensional evolution. In practice, radar-based detection is not only constrained by spatial resolution but also affected by Earth curvature and atmospheric refraction. At longer ranges, even the lowest elevation scans may fail to capture near-surface information, making it challenging to determine whether a mesocyclone will extend to the ground and produce a tornado. In this context, in addition to long-range S-band radars, the incorporation of X-band radars to fill observational gaps is particularly important, especially as phased-array radars offer significant advantages in both spatial and temporal resolution. In regions such as Jiangsu and Guangzhou, radar networks integrating large-scale radars with X-band phased-array systems have already been deployed. Future efforts should further emphasize multi-source data fusion and collaborative multi-radar networks, promoting the development of intelligent, grid-based, and high-precision tornado identification systems, thereby providing stronger support for operational forecasting.

Author Contributions: Conceptualization, G.Z., Q.Z. and F.Z.; methodology, G.Z., Q.Z. and F.Z.; software, G.Z., Q.Z. and F.Z.; validation, G.Z., Q.Z., F.Z. and H.W.; formal analysis, G.Z., F.Z. and T.Y.; investigation, Q.Z. and F.Z.; resources, Q.Z. and F.Z.; data curation, Q.Z.; writing—original draft preparation, G.Z.; writing—review and editing, G.Z., Q.Z., F.Z. and H.W.; visualization, G.Z., T.Y. and H.W.; supervision, Q.Z. and F.Z.; project administration, Q.Z. and F.Z.; funding acquisition, Q.Z. and F.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the National Natural Science Foundation of China (U2342216 and 42575154), National Natural Science Foundation of Sichuan Province (2026NSFSC0208), China Meteorological Administration Tornado Key Laboratory (TKL202309), Key Laboratory of South China Sea Meteorological Disaster Prevention and Mitigation of Hainan Province (SCSF202503), Key Laboratory of Intelligent Meteorological Observation Technology, CMA (ZNGC2024QN03), Key Laboratory of Atmosphere Sounding, CMA (2023KLAS09M, 2024KLAS04M).

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Acknowledgments: The author used S-band radar tornado information from Jiangsu Province and Foshan Meteorological Bureau, and it was greatly appreciated. The authors also thank the published researchers whose literature contains the information used and cited in this paper.

Conflicts of Interest: The authors declare no conflicts of interest.

https://doi.org/10.3390/rs18081124

References

1. Fan, W.; Yu, X. Characteristics of Spatial–Temporal Distribution of Tornadoes in China. *Meteorol. Mon.* **2015**, *41*, 793–805.
2. Zheng, Y.; Liu, F.; Zhang, H. Advances in tornado research in China. *Meteorol. Mon.* **2021**, *47*, 1319–1335.
3. Yu, X.; Zheng, Y. Research and Operational Progress of Contemporary Severe Convective Weather in China. *Acta Meteorol. Sin.* **2020**, *78*, 391–418.
4. Ćwik, P.; Furtado, J.C.; McPherson, R.A.; Taszarek, M. Major May tornado outbreaks in the United States: Novel multiscale atmospheric patterns identified using maximum covariance analysis. *Atmos. Res.* **2025**, *315*, 107872. [CrossRef]
5. Artekha, S. Some generalizations of the tornado model. *Atmos. Res.* **2025**, *329*, 108541. [CrossRef]
6. Zrnić, D.; Burgess, D.; Hennington, L. Automatic detection of mesocyclonic shear with Doppler radar. *J. Atmos. Ocean. Technol.* **1985**, *2*, 425–438. [CrossRef]
7. Wakimoto, R.M.; Wilson, J.W. Non-supercell tornadoes. *Mon. Weather Rev.* **1989**, *117*, 1113–1140. [CrossRef]
8. Lack, S.A.; Fox, N.I. Development of an automated approach for identifying convective storm type using reflectivity-derived and near-storm environment data. *Atmos. Res.* **2012**, *116*, 67–81. [CrossRef]
9. Ryzhkov, A.V.; Schuur, T.J.; Burgess, D.W.; Zrnic, D.S. Polarimetric tornado detection. *J. Appl. Meteorol.* **2005**, *44*, 557–570. [CrossRef]
10. Bluestein, H.B.; French, M.M.; Tanamachi, R.L.; Frasier, S.; Hardwick, K.; Junyent, F.; Pazmany, A.L. Close-range observations of tornadoes in supercells made with a dual-polarization, X-band, mobile Doppler radar. *Mon. Weather Rev.* **2007**, *135*, 1522–1543. [CrossRef]
11. Kumjian, M.R.; Ryzhkov, A.V. Polarimetric signatures in supercell thunderstorms. *J. Appl. Meteorol. Climatol.* **2008**, *47*, 1940–1961. [CrossRef]
12. Snyder, J.C.; Ryzhkov, A.V. Automated detection of polarimetric tornadic debris signatures using a hydrometeor classification algorithm. *J. Appl. Meteorol. Climatol.* **2015**, *54*, 1861–1870. [CrossRef]
13. Zhang, Y.; Bai, L.Q.; Meng, Z.Y.; Chen, B.H.; Tian, C.C.; Fu, P.L. Rapid-scan and polarimetric phased-array radar observations of a tornado in the Pearl River Estuary. *J. Trop. Meteorol.* **2021**, *27*, 81–86.
14. Wang, Y.; Yu, T.Y. Novel tornado detection using an adaptive neuro-fuzzy system with S-band polarimetric weather radar. *J. Atmos. Ocean. Technol.* **2015**, *32*, 195–208. [CrossRef]
15. Hill, A.J.; Herman, G.R.; Schumacher, R.S. Forecasting severe weather with random forests. *Mon. Weather Rev.* **2020**, *148*, 2135–2161. [CrossRef]
16. Zeng, Q.; Qing, Z.; Zhu, M.; Zhang, F.; Wang, H.; Liu, Y.; Shi, Z.; Yu, Q. Application of random forest algorithm on tornado detection. *Remote Sens.* **2022**, *14*, 4909. [CrossRef]
17. Basalyga, J.N.; Barajas, C.A.; Gobbert, M.K.; Wang, J. Performance benchmarking of parallel hyperparameter tuning for deep learning based tornado predictions. *Big Data Res.* **2021**, *25*, 100212. [CrossRef]
18. Sandmæl, T.N.; Smith, B.R.; Reinhart, A.E.; Schick, I.M.; Ake, M.C.; Madden, J.G.; Steeves, R.B.; Williams, S.S.; Elmore, K.L.; Meyer, T.C. The tornado probability algorithm: A probabilistic machine learning tornadic circulation detection algorithm. *Weather Forecast.* **2023**, *38*, 445–466. [CrossRef]
19. McGuire, M.P.; Moore, T.W. Prediction of tornado days in the United States with deep convolutional neural networks. *Comput. Geosci.* **2022**, *159*, 104990. [CrossRef]
20. Xie, J.; Zhou, K.; Chen, H.; Han, L.; Guan, L.; Wang, M.; Zheng, Y.; Chen, H.; Mao, J. Multi-task learning for tornado identification using Doppler radar data. *Geophys. Res. Lett.* **2024**, *51*, e2024GL108809. [CrossRef]
21. Zeng, Q.; Zhang, G.; Huang, S.; Song, W.; He, J.; Wang, H.; Liu, Y. A novel tornado detection algorithm based on XGBoost. *Remote Sens.* **2025**, *17*, 167. [CrossRef]
22. Zhao, K.; Huang, H.; Wang, M.; Lee, W.C.; Chen, G.; Wen, L.; Wen, J.; Zhang, G.; Xue, M.; Yang, Z.; et al. Recent progress in dual-polarization radar research and applications in China. *Adv. Atmos. Sci.* **2019**, *36*, 961–974. [CrossRef]
23. Veillette, M.S.; Kurdzo, J.M.; Stepanian, P.M.; Cho, J.Y.; Reis, T.; Samsi, S.; McDonald, J.; Chisler, N. A benchmark dataset for Tornado detection and prediction using full-resolution polarimetric weather radar data. *Artif. Intell. Earth Syst.* **2025**, *4*, e240006. [CrossRef]
24. Yang, S.; Yang, H.; Shen, F.; Jian, Z. A Review of Image Data Augmentation for Deep Learning. *J. Softw.* **2025**, *36*, 78–79.
25. Hasanin, T.; Khoshgoftaar, T. The effects of random undersampling with simulated class imbalance for big data. In *2018 IEEE International Conference on Information Reuse and Integration (IRI)*; IEEE: New York, NY, USA, 2018; pp. 70–79.
26. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
27. Liu, Z.; Wang, Y.; Vaidya, S.; Ruehle, F.; Halverson, J.; Soljačić, M.; Hou, T.Y.; Tegmark, M. Kan: Kolmogorov-Arnold networks. *arXiv* **2024**, arXiv:2404.19756.
28. Taud, H.; Mas, J.F. Multilayer perceptron (MLP). In *Geomatic Approaches for Modeling Land Change Scenarios*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 451–455.

https://doi.org/10.3390/rs18081124

29. Gordon, W.J.; Riesenfeld, R.F. B-spline curves and surfaces. In *Computer Aided Geometric Design*; Elsevier: Amsterdam, The Netherlands, 1974; pp. 95–126.
30. Strang, G.; Fix, G.J. *An Analysis of the Finite Element Method*; Wellesley-Cambridge Press: Philadelphia, PA, USA, 2008; p. 402.
31. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

https://doi.org/10.3390/rs18081124

fire

MDPI

Article

Deep Unsupervised Learning for Indoor Fire Detection Using Wi-Fi Signals

Sara Mostofi [1], Fatih Yesevi Okur [1,2,*], Ahmet Can Altunişik [1,2,3] and Ertugrul Taciroğlu [4]

[1] Department of Civil Engineering, Karadeniz Technical University, Trabzon 61080, Türkiye; saramostofi.4e@gmail.com (S.M.); ahmetcan@ktu.edu.tr (A.C.A.)
[2] Earthquake and Structural Health Monitoring Research Center, Karadeniz Technical University, Trabzon 61080, Türkiye
[3] Dynamic Academy Software Construction Industry Trade Limited Company, Trabzon 61080, Türkiye
[4] Civil and Environmental Engineering Department, University of California Los Angeles, Los Angeles, CA 90095, USA; etacir@ucla.edu
* Correspondence: yesevi@ktu.edu.tr

Abstract

This study proposes a sensor-free approach for indoor fire detection that leverages existing Wi-Fi infrastructure as a passive sensing modality. By extracting Channel State Information (CSI) from prevalent 802.11n Wi-Fi signals and applying an unsupervised deep anomaly detection model, the approach conceptualizes the wireless environment as a sensorless detection field capable of identifying combustion-induced perturbations without requiring any physical sensors. CSI data were collected in both normal and flame-induced states under three combustion conditions (gasoline, wood, plastic), each introducing unique signal perturbations. These data, which exhibit diverse signal perturbations, were used as input to four unsupervised deep anomaly detection architectures: a variational autoencoder (VAE), a 1D convolutional autoencoder (CNN-AE), a long short-term memory autoencoder (LSTM-AE), and a hybrid CNN-LSTM autoencoder. Each architecture was trained exclusively on baseline data to learn compact latent representations of normal signal patterns. Among the evaluated architectures, CNN-AE achieved perfect detection across all scenarios, showing high responsiveness to signal disruptions. LSTM-AE tracks prolonged combustion but struggles with fast-onset anomalies. VAE maintains low error during baseline but misses sharp deviations. These findings validate that Wi-Fi CSI encodes latent combustion features. The method requires no additional sensors and operates on existing signals, enabling scalable smart building integration via lightweight software updates.

Keywords: fire detection; Channel State Information (CSI); Wi-Fi; unsupervised deep anomaly detection; autoencoders; CNN-LSTM

1. Introduction

Residential fire safety is a cornerstone of urban sustainability, touching on issues of human safety, infrastructure durability, and economic stability [1,2]. The impacts of residential fires are far-reaching, including loss of life, damage to infrastructure, and long-term financial strain [3,4]. Central to mitigating these consequences is the early detection of ignition and early combustion stages. Early detection governs evacuation timing, suppression system engagement, and emergency service mobilization [5]. A shorter detection latency substantially increases the likelihood of containment and life preservation.

As such, recent efforts in fire protection are directed toward minimizing the time lag between fire onset and alarm activation [6–9].

Residential fire detection generally relies on devices designed to identify early combustion indicators based on heat or smoke [10,11]. While their performance is validated and their usage widespread, they operate on threshold-based activation models that may not pick up the very earliest signs of combustion, sometimes delaying an alarm or triggering a false alarm [12,13]. Moreover, despite being a foundational component of fire protection, fire detection infrastructure remains alarmingly deficient across residential properties. A considerable proportion of residences either lacks detection systems or contains devices that are inoperative or fail to meet minimum standards, posing significant risks to occupants and responders alike [4]. Data from the U.S. Fire Administration indicate that smoke alarms were present in just 33% of fatal residential fires, absent in 25%, and unconfirmed in 42% of cases. Within the subset of homes where alarms were confirmed, only 16% functioned correctly during the fire event. Alarm failure occurred in 7% of those cases, while 11% had unknown operational status. Notably, incidents where fires were too minor to activate the alarm comprised less than 1% of failures [14]. These figures underscore systemic limitations in both the deployment and performance fidelity of residential fire alarm systems.

Although fire codes stipulate alarm installations across residential types from single-family units to multi-dwelling structures, real-world enforcement and device operability are highly variable and difficult to ensure. For larger complexes, requirements extend to centralized, interconnected systems [15]. However, enforcement is especially inconsistent in aging or non-retrofitted dwellings. Older residential buildings, which constitute a substantial fraction of global housing, often face architectural constraints that hinder effective fire detection [16]. Narrow corridors, closed ceilings, and preservation laws make alarm installation disruptive or expensive. Moreover, features like high ceilings and enclosed layouts distort smoke pathways, delaying detector activation [17].

Advancements in video-based fire detection have expanded the field of contactless fire monitoring [18,19]. Vision-based solutions, often integrated into security cameras, utilize deep learning to detect visual signatures of fire, such as smoke trails or flame oscillations [19–22]. These are particularly effective in expansive or tall spaces where conventional detectors may be delayed. Remote visual confirmation also enhances situational awareness for emergency response. However, vision models demand continuous illumination, unobstructed views, and significant computational resources, often processed on edge or cloud platforms [23–25].

Modern frameworks emphasize interoperability, integrating fire detection with existing systems, including HVAC management, and automated suppression systems [26–28]. These frameworks employ internet connectivity to link multiple detectors, such as smoke, heat, or multi-sensor, each uniquely identified within the network [6,29,30]. Ming [31] introduced a smart fire intervention suite with optical sensors, all coordinated via Wi-Fi. Ehsan et al. and Lakshmi et al. [9,32] developed wireless fire alarm networks with mobile notification capabilities.

A novel direction in fire detection utilizes ambient Wi-Fi signals as non-contact sensors. The application of Wi-Fi signal analysis has seen significant growth in various fields beyond communications, most notably in security, healthcare, human–computer interaction, and occupancy monitoring. Channel State Information (CSI), derived from pervasive Wi-Fi signals, has demonstrated capabilities in motion tracking [33], human identification [34], and respiratory monitoring [34–36]. Its widespread applicability stems from the ubiquity of Wi-Fi and its cost-efficient infrastructure.

https://doi.org/10.3390/fire9050189

In construction, Wi-Fi-enabled systems have addressed safety through unauthorized access monitoring [37] and contactless fall detection via deep learning [38]. Material analysis has similarly benefited, with Convolutional Neural Network (CNN) processing CSI magnitude and phase to identify material thickness, compaction, and moisture content with very high accuracy [39,40]. Studies have confirmed the capabilities of Wi-Fi detection systems for structural health monitoring [41]. Advances in hardware and embedded computing have enabled the deployment of low-cost structural monitoring solutions based on single-board computers such as Raspberry Pi. These platforms support continuous, high-resolution monitoring without degrading data fidelity, representing a practical substitute for high-cost proprietary systems [42].

In parallel, Wi-Fi sensing is being incorporated into fire detection systems to provide efficient, infrastructure-wide hazard alerts. Wi-Fi-based fire detection employs Radio Frequency (RF) emissions from Wi-Fi transmitters as probes of the surrounding environment [43,44]. It leverages changes in CSI, which encapsulates amplitude and phase variations in RF propagation, to infer environmental perturbations caused by fire. These systems offer rapid, non-invasive installation, an advantage particularly relevant for complex or old structures where physical rewiring may be impractical or prohibited. Integration with smart home ecosystems adds convenience. The inherent modularity of Wi-Fi-based platforms allows for seamless expansion and relocation of detection units, avoiding structural disruption. Battery-powered configurations further ensure operational continuity during electrical outages, addressing a critical reliability criterion. Moreover, such systems are well-suited to cost-constrained or spatially restricted environments [45,46], broadening access to high-quality fire monitoring.

Despite its ubiquity, research exploring the utility of Wi-Fi infrastructure for fire detection in residential settings remains in its early stages. The Wi-Fire system by [43] marked one of the earliest demonstrations of passive, device-free fire detection. Leveraging 802.11n hardware, the system extracted thermal signal features through PCA-reduced, denoised CSI streams and employed a Random Forest classifier to detect fire events. Subsequent work by [47] advanced this field by demonstrating temperature profiling through CSI acquired from Raspberry Pi 4B devices, revealing that CSI amplitude tightly correlates with air temperature. Subsequent studies expanded into real combustion scenarios, with CSI collected at up to 1350 fps on both 2.4 and 5 GHz bands. Using Raspberry Pi platforms to sample CSI at high frame rates (up to 1350 fps) across 2.4 and 5 GHz bands, the system recorded RF behavior over flame exposure. By analyzing changes in CSI amplitude and variance, the authors identified distinct amplitude patterns for fire and smoke stages. Their findings support the potential of CSI as a low-cost fire detection tool [44]. In another track [48], ESP32 transceivers were used in a constrained metal-box configuration to localize heat sources with CSI captured at 20 Hz over 64 subcarriers. The top 40 high-variance subcarriers were vectorized and classified using a linear support vector machine to determine active heater positions.

Although current research has contributed to developing Wi-Fi CSI-based fire detection systems, key challenges must be addressed for effective real-world implementation. Most studies have relied on narrowly defined ignition sources, typically flame sources with limited smoke production or thermal variability. These fail to represent the RF-interaction complexity posed by diverse fuel classes such as synthetic polymers or biomass, all of which emit distinct combustion by-products that influence CSI differently. Moreover, many of the current detection systems rely on supervised learning frameworks or fixed-threshold classifiers, which introduce rigidity. These models exhibit lower resilience to untrained fire scenarios and are not fully equipped for dynamic environments. Addressing the challenges of current fire detection systems [49], this study proposes a more robust and scalable solu-

https://doi.org/10.3390/fire9050189

tion via unsupervised anomaly detection using autoencoder networks operating on CSI streams from ESP32 Wi-Fi modules. Four deep learning architectures, namely Variational Autoencoder (VAE), Convolutional Autoencoder (CNN-AE), Long Short-Term Memory Autoencoder (LSTM-AE), and a hybrid CNN–LSTM model, are explored due to their proven capabilities in encoding temporal, spatial, and statistical aspects of time-series data [50–53]. Each model learns to compress normal signal behavior into a latent space and reconstruct input data with minimal loss [54].

The VAE adopts a probabilistic generative framework that learns latent variable distributions, enabling principled anomaly detection through reconstruction likelihoods [55]. The 1D-CNN is optimized for extracting localized spatial regularities within individual time frames, capturing granular sub-patterns. The LSTM-AE, with its recurrent design, captures inter-window temporal structure and long-range dependencies [55,56]. The hybrid CNN–LSTM structure unifies local feature extraction and sequence modeling into a layered hierarchy [53,57], effectively capturing anomalies during the ignition phase, where perturbations are minimal but meaningful. Given the uniquely subtle distortions induced by thermal gradients and scattering effects in the ignition phase [58], the performance of these four architectures is compared to select the most responsive model for early-stage fire detection in indoor environments. The end-to-end pipeline for the CSI-based early fire detection system is illustrated in Figure 1.

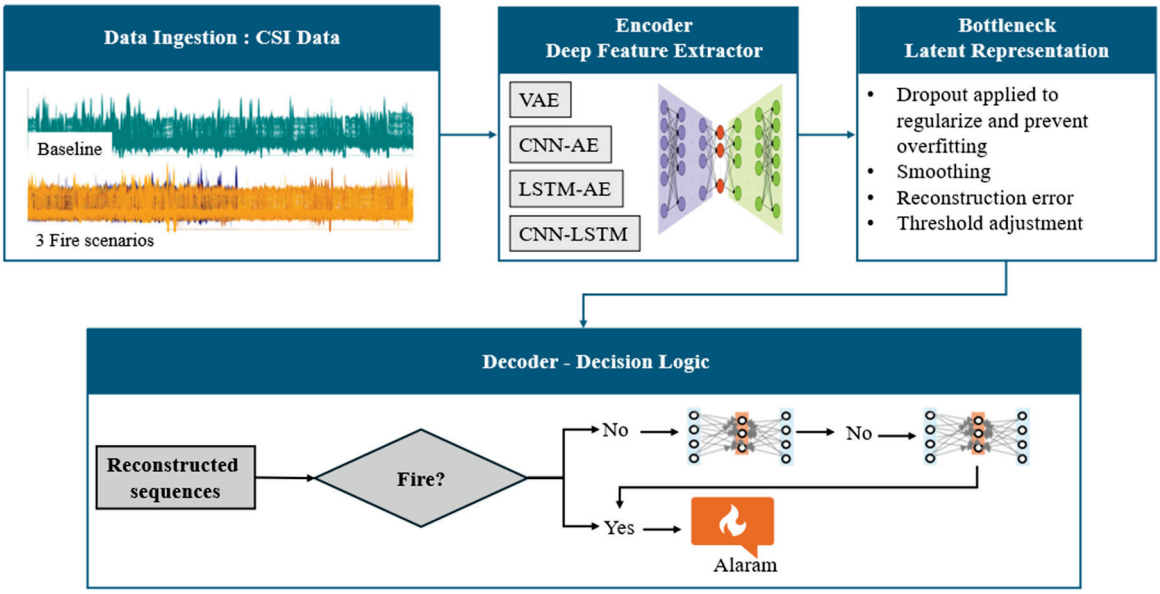


Figure 1. The proposed early-stage fire detection system.

As shown in Figure 1, the proposed pipeline for unsupervised anomaly detection begins with CSI signal ingestion, wherein raw CSI sequences are captured under both normal and fire-affected conditions across three distinct scenarios. These high-dimensional sequences are processed using deep autoencoder networks with bottleneck layers to extract compressed latent patterns. During training, reconstruction loss is minimized exclusively on fire-free data, learning the characteristics of the signal under normal conditions. Dropout, smoothing, and adaptive thresholding are incorporated to regularize the training process, suppress transient noise, and dynamically adjust detection sensitivity. Using the CSI data collected under combustion conditions, subtle differences in signal behavior were identified and used to calibrate thresholds. These thresholds were adaptively adjusted to reflect the degree of deviation from the learned normal pattern, thereby improving detection precision under different ignition scenarios. When deployed, the decoder compares each

https://doi.org/10.3390/fire9050189

reconstructed sequence with its original counterpart, and the resulting error becomes the key signal for anomaly detection [59,60]. If this reconstruction loss surpasses a threshold across multiple time windows, the input is flagged as anomalous, signaling a combustion event. Based on this analysis, the model demonstrating optimal sensitivity to early fire signatures and superior robustness is selected for the early fire detection alarm system.

2. Methodology

The methodology adopted in this study followed a three-stage pipeline: CSI data acquisition, signal preprocessing, and anomaly-based fire detection using unsupervised deep learning models, as illustrated in Figure 2.

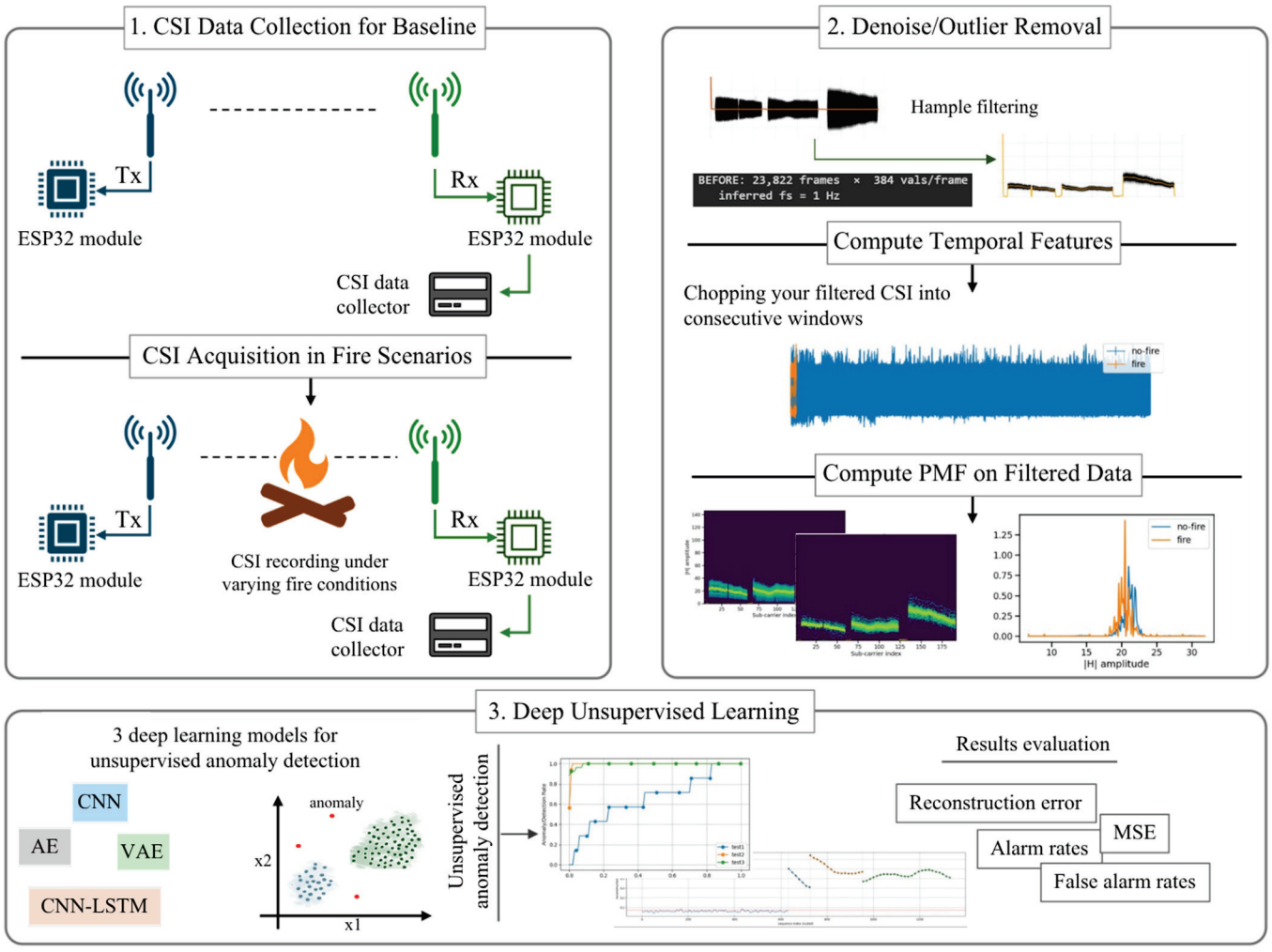


Figure 2. Overview of the study.

In the first stage, CSI data were collected using two ESP32 modules configured as transmitter (Tx) and receiver (Rx), positioned approximately 2.0 m apart under line-of-sight conditions. Both nodes were elevated 0.5–1.0 m above the floor, and the combustion source was placed near the midpoint of the direct propagation path between them. CSI streams were recorded under two settings:

- Baseline conditions: collecting CSI under normal conditions with no fire activity, allowing models to learn the normal propagation characteristics of indoor wireless signals.
- Fire scenarios: recording RF behavior during three controlled combustion events involving gasoline, wood, and plastic that were introduced between the devices to simulate real-world ignition phases.

https://doi.org/10.3390/fire9050189

In the second stage, the raw CSI signals were preprocessed to reduce noise and enhance interpretability. A Hampel filter was used to eliminate outliers caused by hardware or environmental jitter. The cleaned signals were segmented into overlapping time windows, and statistical features, such as mean, standard deviation, and skewness, were computed for each segment. This enabled localized analysis of signal behavior and prepared the data for input into the detection models.

In the final stage, four deep learning models, namely VAE, CNN-AE, LSTM-AE, and a hybrid CNN-LSTM, were trained on fire-free baseline data. These models learned the normal patterns of wireless signal propagation and were tested on fire data to identify anomalies through reconstruction errors. Detection thresholds were dynamically set based on baseline error distributions, enabling the system to flag deviations likely caused by combustion. Each model's performance was evaluated in terms of detection accuracy, false alarm rate, and robustness across different fire types.

2.1. Data Acquisition and Experimental Setup

The CSI sensing system was configured using two ESP32 microcontrollers and a LattePanda operating on Windows 11. The ESP32 was chosen for its built-in CSI acquisition capabilities, low cost, dual-core 240 MHz processor, and integrated Wi-Fi/BLE functionality, which support both data transmission and on-board inference. It allows access to CSI, which provides subcarrier-level information of amplitude and phase, eliminating the need for specialized hardware to access CSI and allowing real-time data processing or transmission [61]. This helps in capturing the subtle variations in RF propagation induced by thermal gradients and soot particles, and refractive-index fluctuations during combustion events.

The experimental setup, as shown in Figure 3, consists of a Wi-Fi transmitter and receiver placed within a room such that a potential fire source lies in the propagation path between them. Both devices were raised to approximately 0.5–1 m above the ground to minimize ground-induced multipath. One ESP32 device operates as a Wi-Fi transmitter while another ESP32 serves as the CSI receiver. The ESP32 receiver captures CSI for each received Wi-Fi frame by measuring the complex-valued Channel Frequency Response (CFR) across all Orthogonal Frequency Division Multiplexing (OFDM) subcarriers [62]. Formally, the CFR at frequency f and time t, denoted as H(f,t), can be expressed as Equation (1) [44]:

$$H(f,t) = H_s(f,t) + \sum_{k\epsilon P} ak(f) exp(-j\frac{2\pi(d_k(0) + v_k t)}{\lambda}) \quad (1)$$

Here, $H(f,t)$ denotes the complex CFR at frequency f and time t, and $H_s(f,t)$ denotes its static component due to stationary reflectors. P is the set of dynamic propagation paths. For each path $k \in P$, $a_k(f)$ denotes the complex amplitude, $d_k(0)$ the initial propagation-path length at $t = 0$, and v_k the effective rate of path-length change. Finally, λ denotes the carrier wavelength and $j = \sqrt{-1}$ the imaginary unit. Combustion perturbs the propagation channel through heat, smoke, and refractive-index fluctuations, which alter the dynamic multipath terms and hence the measured CSI.

Fire affects CSI by modifying the RF propagation environment through several coupled physical mechanisms. The heat released during combustion creates strong local temperature and density gradients in air, which alter its refractive index and perturb the propagation paths between the transmitter and receiver. At the same time, smoke particles and combustion by-products introduce additional attenuation and scattering, while the resulting turbulent plume continuously reshapes the multipath structure of the channel. These effects cause measurable changes in CSI amplitude and phase by altering path loss, reflection strength, and path stability [43,63]. Different fuels generate distinct signa-

https://doi.org/10.3390/fire9050189

tures due to variations in heat release and smoke properties, supporting fuel-dependent anomaly detection.

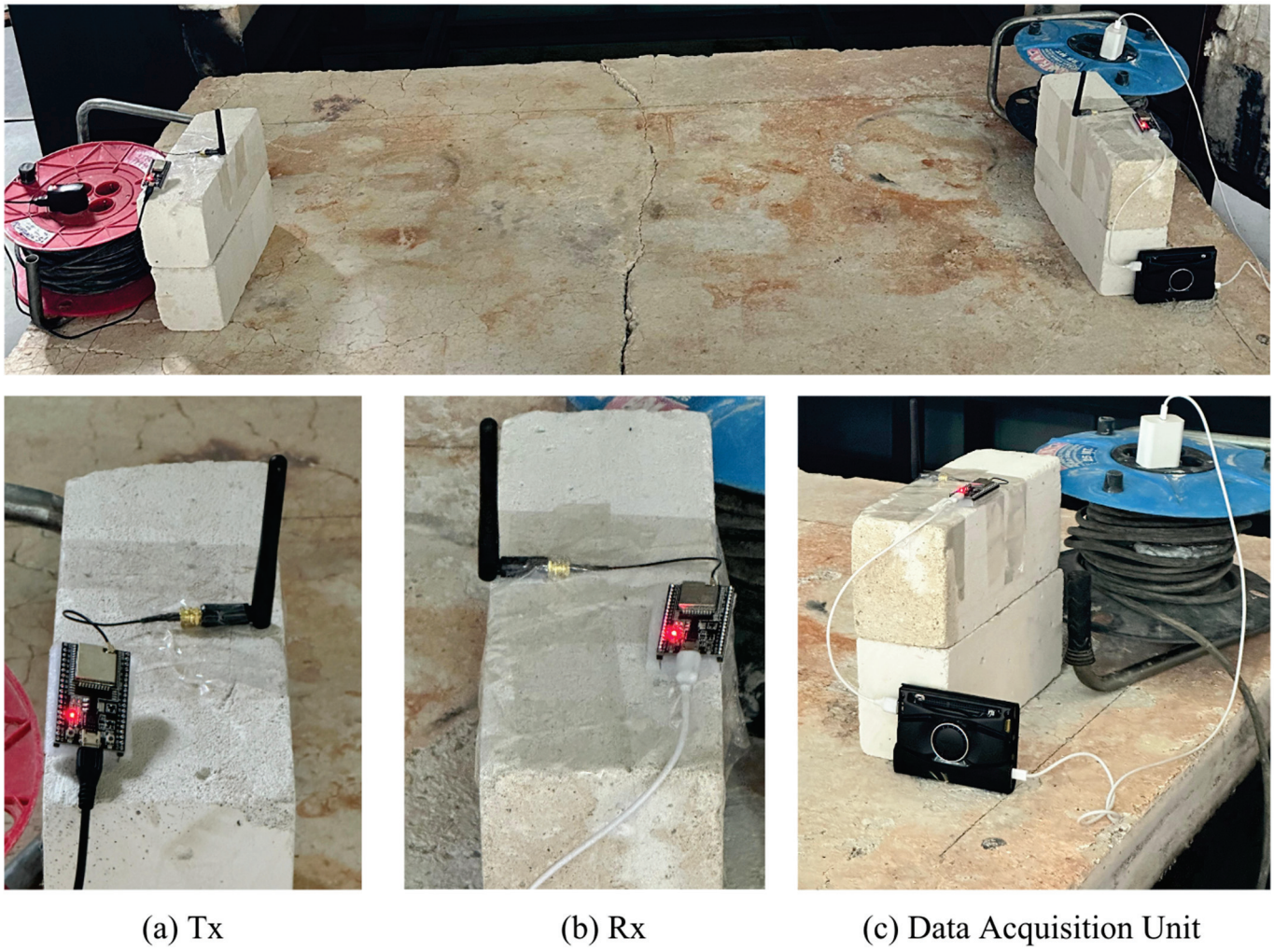


Figure 3. Experimental setup.

Each CSI sample, per the 802.11n specification, comprises amplitude and phase readings for subcarriers [62]. The analysis in this work prioritizes amplitude data because it provides a more stable and direct indication of fire-induced scattering and attenuation effects in the wireless channel, while avoiding the additional calibration complexity associated with phase data [43,64]. The system incorporated no additional sensing units, relying exclusively on CSI readings from existing Wi-Fi signals. CSI data collection was divided into two phases: baseline and fire event recordings. For baseline data, the environment was monitored under normal conditions (no flame or smoke) for an extended period of 3 h to capture the typical CSI variability from benign environmental changes and hardware noise. This baseline established the "normal" CSI profile for the area.

Subsequently, three controlled fire events were introduced between the ESP32 transmitter and receiver: (1) a gasoline flame for 10 min, (2) a smoldering wood fire for 5 min, and (3) a plastic fire for 10 min (Figure 4). The differing combustion types were selected to induce varying environmental effects such as thermal intensity, particulates, and smoke, and to capture a diverse range of fire-induced signal disturbances. The experimental model assumed the absence of human occupants, representing conditions under which the fire alarm system must operate autonomously to detect an active flame. Each experiment provided a time series of CSI amplitude and phase across subcarriers.

https://doi.org/10.3390/fire9050189

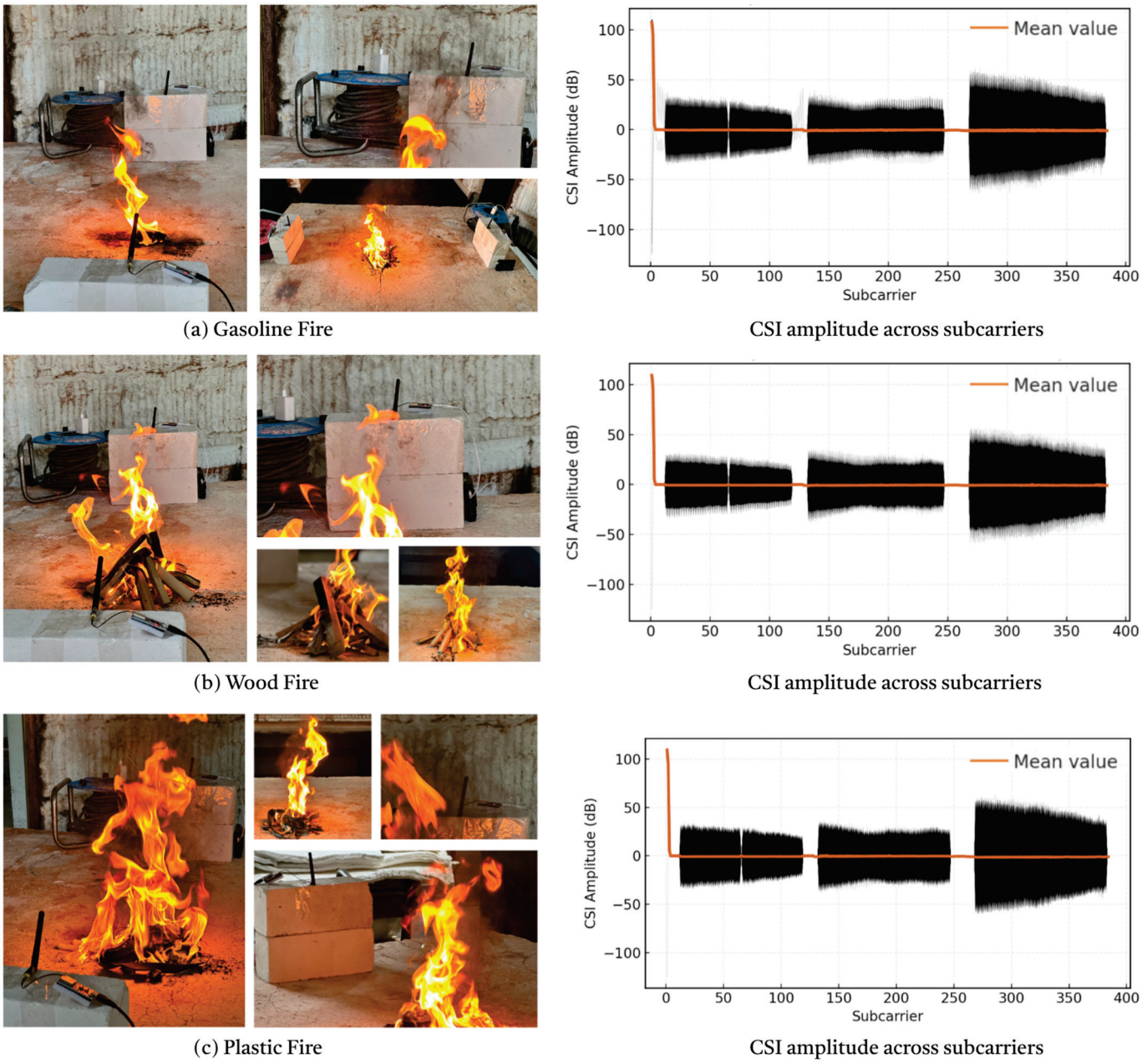


Figure 4. Fire tests and CSI acquisition.

Figure 5 provides a dense visual encoding of combustion-induced RF perturbations, contrasting baseline CSI profiles with those measured during different fire conditions. Unprocessed CSI amplitude readings offered detailed insight into the physical effects of fire on wireless propagation. These raw packets preserved the true per-packet fading signature of the environment. Each vertically offset trace illustrates the subcarrier profile of a single packet, highlighting deviations in amplitude patterns attributable to flame propagation and smoke dynamics.

Across all tested fire types, raw CSI traces revealed how RF propagation responded to combustion properties. The rapid and localized gasoline flame primarily reduced the amplitude of higher-index subcarriers, with minimal overall impact. Wood combustion caused more consistent amplitude suppression across the spectrum and introduced jitter at lower frequencies, likely due to sustained smoke and heat. The plastic fire produced the broadest and most severe signal degradation, consistent with heavy soot and material off-gassing. These unprocessed overlays served as visual indicators of the combustion-

https://doi.org/10.3390/fire9050189

induced channel transformations, validating that wood and plastic fires produce clearer spectral signatures within CSI data. The results emphasized that distinct fire types impose materially different propagation dynamics. They underscored the advantage of CSI in identifying broadband fires like plastic and wood, making it promising for early fire detection in residential buildings.

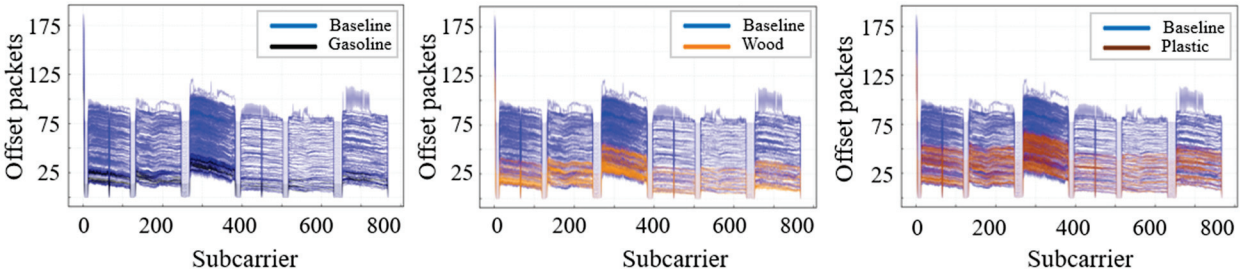


Figure 5. Pure packet values across baseline and fire conditions.

2.2. Preprocessing

The CSI time series underwent preprocessing to reduce noise while maintaining subtle signal deviations associated with combustion events. The initial phase involved application of the Hampel filter, a non-parametric technique recognized for its robustness in outlier detection within Wi-Fi sensing literature [43,44,47]. The Hampel filter is particularly suited for denoising CSI amplitude signals, where short-duration spikes often arise from hardware jitters or RF noise, rather than environmental changes. This method preserves sustained trends while removing impulsive outliers [65]. This distinction is vital in fire detection tasks, where combustion-induced anomalies appear as gradual, low-frequency signal drifts rather than instantaneous transitions.

The Hampel filter operates over a sliding temporal window, where each data point is evaluated against the local median and flagged as an outlier if it deviates by more than a predefined threshold, typically three times the scaled median absolute deviation (MAD). These outliers are substituted with the median value of the corresponding window, thereby mitigating high-frequency noise without disrupting the long-term signal structure [66,67] This method is well-suited for time-series sensor data as it can preserve sustained trends while removing impulsive outliers [67]. Hence, Hampel filtering was used to obtain an interpretable CSI amplitude trace. Once denoised, the CSI data stream is partitioned into fixed-duration windows, enabling localized analysis of temporal dynamics and standardized input to deep learning models. The sampling frequency was set to 1 Hz. Segmentation captured the temporal dynamics of the signal in discrete and manageable segments and standardized the input dimensions for deep autoencoder models. The obtained mean CSI amplitudes after denoising are shown in Figure 6.

To capture statistical irregularities induced by combustion, the probability mass function (PMF) was calculated for CSI amplitude values in each temporal segment. The core premise is that fire-related activity causes measurable shifts in these distributions, driven by changes in the physical channel, especially from thermal scattering, absorption, and dynamic multipath behavior. In higher-temperature fires, ionization may occur, forming localized plasma that disturbs the RF channel further [43]. Figure 7 illustrates the environmental dynamics encoded by the PMF, shift, and spread of CSI responses across subcarriers.

https://doi.org/10.3390/fire9050189

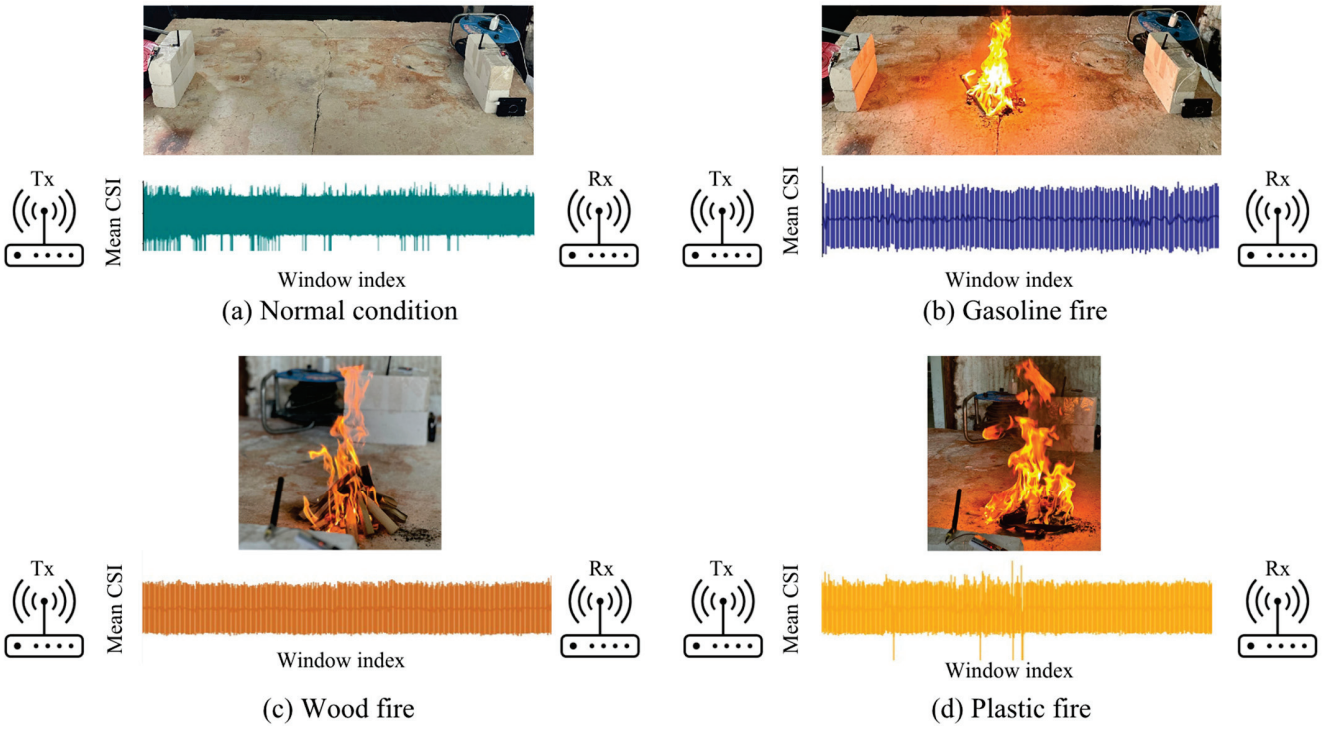


Figure 6. Denoised CSI amplitudes.

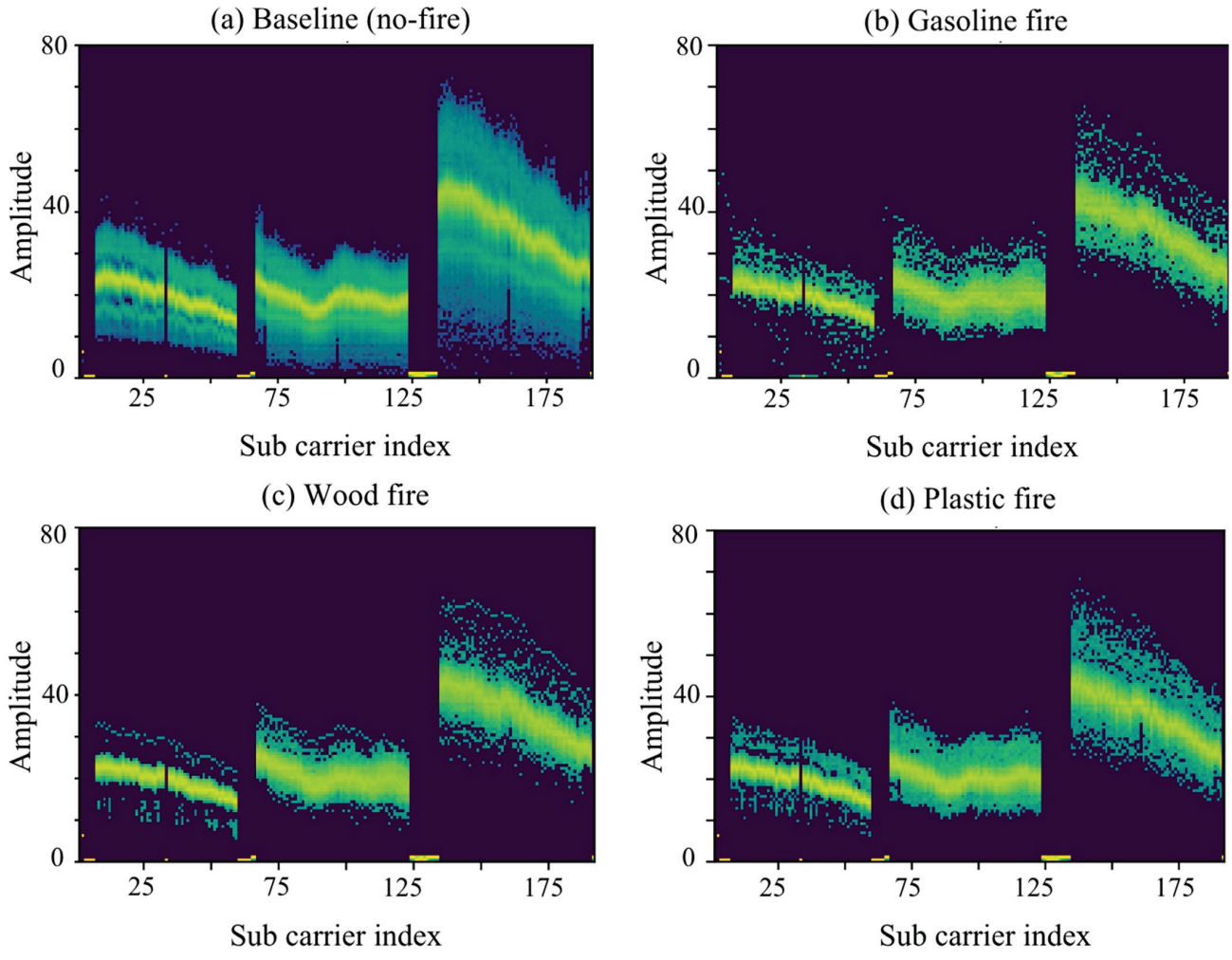


Figure 7. PMF-based CSI amplitude shifts under different conditions.

https://doi.org/10.3390/fire9050189

In baseline conditions, the CSI amplitude distributions exhibited compact, unimodal PMFs with minimal variance, indicative of a stable, undisturbed multipath field. However, combustion events introduced substantial spectral deviations. Gasoline broadened the PMF at higher-index subcarriers, reflecting signal degradation from short bursts of combustion gases. Wood fire distributions broadened gradually, demonstrating slow-forming thermal turbulence. Dense smoke and high thermal loads from plastic resulted in more widespread, chaotic PMF structures, which can serve as strong indicators for anomaly detection. Figure 8 shows four subcarriers whose amplitude PMFs undergo notable shifts under combustion. These subcarriers were selected based on the degree of amplitude distribution divergence between the no-fire and three fire conditions, independent of time order. PMF analysis was used for exploratory interpretation of combustion-induced CSI distribution shifts and was not directly used as model input.

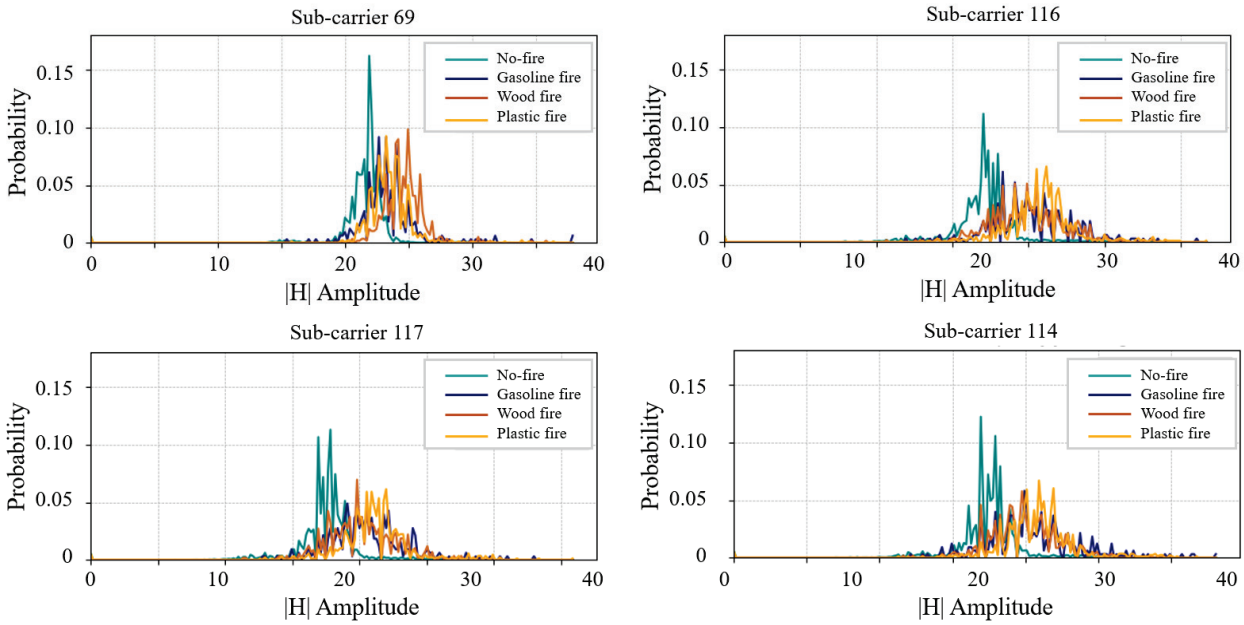


Figure 8. Top amplitude distribution shifts across no-fire and fire events.

In addition, a set of summary statistical features from each window of CSI data was extracted. These included the mean (capturing overall attenuation), standard deviation, interquartile range (IQR) (reflecting spread and signal volatility), skewness and kurtosis (encoding distribution asymmetry and tail behavior), as well as the median and range. Each of these metrics maps to a structural feature of the RF signal, enabling interpretable, low-dimensional embeddings of complex channel dynamics [68]. To encode a stable and informative representation of the wireless channel, CSI amplitudes were temporally averaged across all subcarriers at each timestamp. This leveraged the correlation across tightly spaced subcarriers in 802.11n, reducing dimensionality while retaining global multipath behavior.

2.3. Deep Unsupervised Learning Models

To detect fire-induced anomalies from indoor wireless signals, four unsupervised deep learning architectures including VAE, CNN-AE, LSTM-AE, and CNN-LSTM were implemented. A key technique underlying these models is autoencoder-based anomaly detection, where the model is trained exclusively on normal (non-fire) CSI inputs to accurately reconstruct typical signal patterns. Autoencoder-based anomaly detection was used

https://doi.org/10.3390/fire9050189

to learn the normal CSI pattern. For an input sample x_i, the encoder $h_E(\cdot;\theta_E)$ maps the input to a latent representation z_i, and the decoder $h_D(\cdot;\theta_D)$ reconstructs the input as $\hat{x}_i$:

$$z_i = h_E(x_i;\theta_E), \hat{x}_i = h_D(z_i;\theta_D) \tag{2}$$

The reconstruction loss used for training the deterministic autoencoder models is defined as Equation (3) [69]:

$$\mathcal{L}_{\text{rec}}(\theta_E,\theta_D) = \frac{1}{N}\sum_{i=1}^{N} \parallel x_i - \hat{x}_i \parallel_2^2 \tag{3}$$

where N is the number of training samples. The encoder and decoder parameters are optimized by minimizing $\mathcal{L}_{\text{rec}}$ on fire-free baseline data only.

In practical deployment, when CSI patterns are perturbed by combustion, the reconstruction loss increases. To find the best configuration, a 20-epoch pilot grid search explored combinations of hidden sizes, latent dimensions, and learning rates. Inputs exceeding a predefined threshold were classified as fire-related anomalies. During inference, anomalies were detected by computing the mean squared error (MSE) between the original input and its reconstruction. A statistical threshold, based on a 95% confidence level, separated nominal deviations from true anomalies. During inference, reconstruction errors were smoothed using a moving average, and anomaly scores were thresholded at $\mu + 1.96\ \sigma$ of the training reconstruction MSE.

The VAE model was designed to capture probabilistic latent structures in CSI time series. During encoding, input windows were flattened and passed through dense layers that output μ and $\log \sigma^2$. The decoder mirrored this architecture, expanding latent vectors back into the original CSI shape. The model leveraged the reparameterization trick to ensure smooth latent interpolation and prevent overfitting. The decoder reconstructed the original input shape symmetrically by expanding the latent code through mirrored fully connected layers. All input features were standardized based on statistics computed from the initial 80% of a baseline dataset, while the remaining 20% was allocated for validation. Additionally, three independent traces were withheld for final testing. A two-stage grid search was conducted across hidden layer sizes {16, 32, 64}, latent dimensionalities [2–32] and learning rates $\{1 \times 10^{-3}, 1 \times 10^{-4}\}$. The final model utilized 64 hidden units, a 17-dimensional latent space, and a learning rate of 0.001. This configuration was trained for 200 epochs with a constant KL-divergence weight of 1.0 and 20% dropout.

The CNN-AE consisted of convolutional layers that act as an encoder, compressing the input into a lower-dimensional latent feature map, followed by deconvolution layers as a decoder that reconstructs the input matrix. The encoder consisted of two convolutional layers (kernel size 3, padding 1), reducing the input down to latent feature map depths of 16, 32, or 64. These layers were interleaved with ReLU activations and 20% dropout to encourage robustness. A mirrored decoder architecture, using transpose convolutions, reconstructed the original input sequence. Sliding windows were constructed identically to prior models. A grid search over parameters similar to VAE guided architectural selection. The final configuration included 64 hidden units, a latent space dimensionality of 17, and a 0.001 learning rate.

The LSTM-AE model was structured around a recurrent architecture. The encoder processed each temporal window through an LSTM layer, after which the final hidden state was projected into a latent vector via a fully connected layer. Consistent with the grid search parameters utilized for the VAE, the latent representation was optimized over combinations of hidden layer sizes, latent dimensionalities, and learning rates. During

https://doi.org/10.3390/fire9050189

decoding, this latent vector was replicated across time steps and passed through a second LSTM decoder, culminating in a final linear layer that reconstructed the feature values for each time step. Training followed the same 80/20 data partitioning, normalization on training data, and windowing strategy as previous models. The final model was trained using a 17-dimensional latent space, 64 hidden units, and a learning rate of 0.001.

The CNN–LSTM autoencoder integrated a convolutional encoder for subcarrier feature extraction and an LSTM encoder for sequential modeling. Two convolutional layers with a kernel size of 3 and padding of 1, each followed by a ReLU, acted as the spatial encoder. The resulting map was transposed and fed into an LSTM encoder to capture the sequential dynamics across windows of length 5. The LSTM decoder outputs were then reshaped and passed through the convolutional decoder, a mirrored architecture of the encoder. The model, optimized using Adam with 1×10^{-3} weight decay, achieved stable training with 64 hidden units, 32 latent dimensions, and a learning rate of 0.001.

3. Results and Discussion

Post-training evaluation aimed to assess how well the models generalized to unseen sequences and captured fire-induced signal deviations. The analysis included reconstruction-error patterns (Figure 9), MSE over 200 epochs (Figure 10), and formal anomaly detection metrics, indicative of model utility under deployment conditions, as compared in Table 1. Figure 9 illustrates the smoothed reconstruction error signals for each model, contrasting their behavior under baseline (training/validation) conditions with test sequences involving different fire conditions. The horizontal dashed line represents the data-driven threshold defined as Equation (4):

$$\tau = \mu_{train} + 1.96\,\sigma_{train} \tag{4}$$

Table 1. Detection comparison across all models.

Model	Test 1: Gasoline	Test 2: Wood	Test 3: Plastic	FAR (%)	OAR (%)	ARL
VAE	14.3	100.0	96.2	4.0	86.0	23.8
CNN-AE	100.0	100.0	100.0	3.6	100.0	24.3
LSTM-AE	71.4	100.0	96.2	4.9	94.0	20.8
CNN-LSTM	28.6	100.0	96.2	5.5	88.0	16.9

As demonstrated in Figure 9a, VAE effectively captured wood and plastic fires but exhibited insufficient sensitivity to gasoline-based perturbations. The higher reconstruction noise floor and smoothed latent representation likely contributed to this limitation. Only transient error peaks were observed, resulting in a 14.3% detection rate in that context due to subtle error fluctuations. The CNN-AE maintained a tightly bound baseline error well below its τ, and all fire scenarios produced clear and sustained error excursions above this threshold. This consistent pattern enabled perfect detection across all fire types. LSTM-AE was similarly effective for wood and plastic sequences but missed a small part of gasoline-related anomalies. This can be attributed to its recurrent structure, which smoothed rapid changes and reduced responsiveness to short-lived spikes.

https://doi.org/10.3390/fire9050189

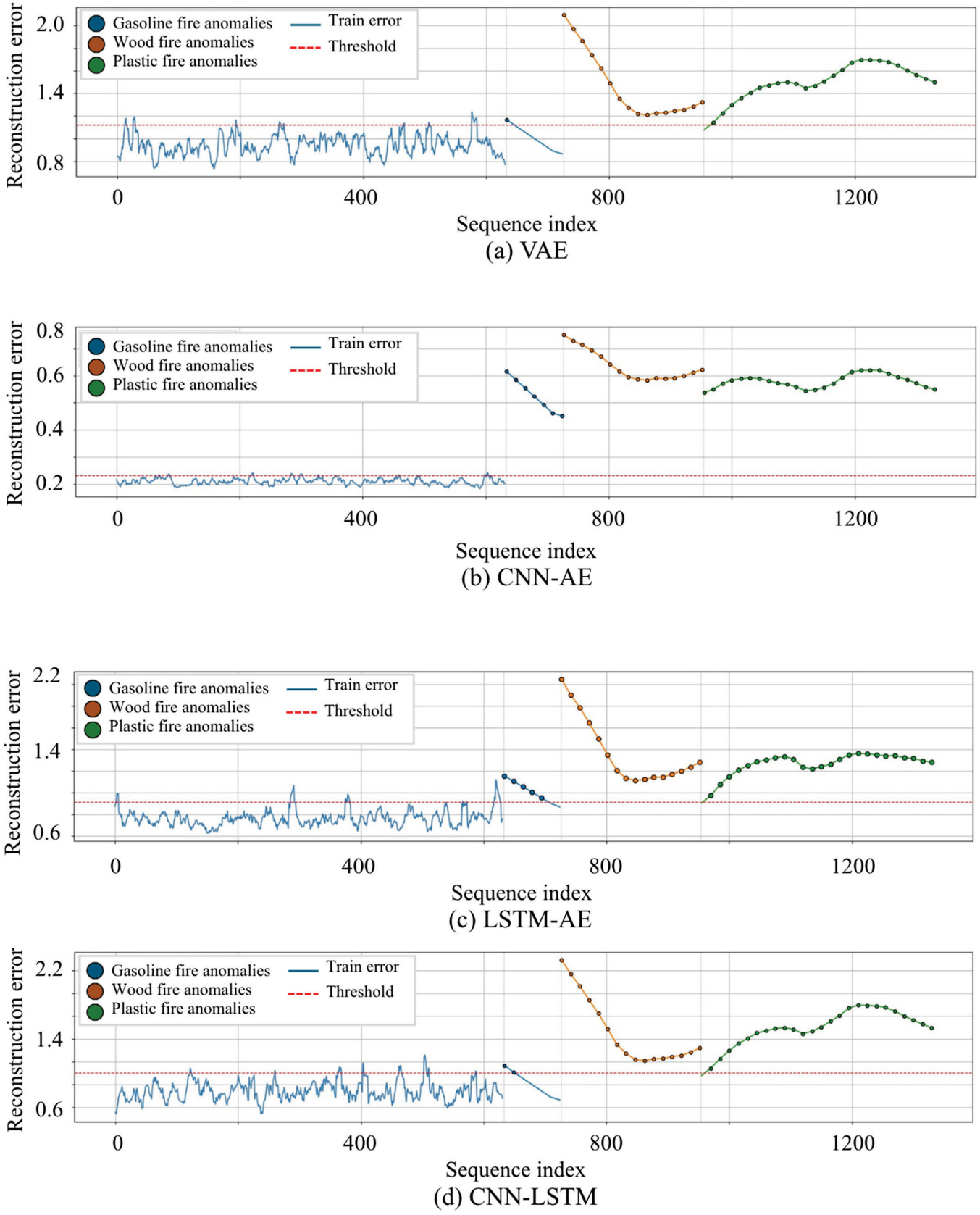


Figure 9. Reconstruction error timelines.

https://doi.org/10.3390/fire9050189

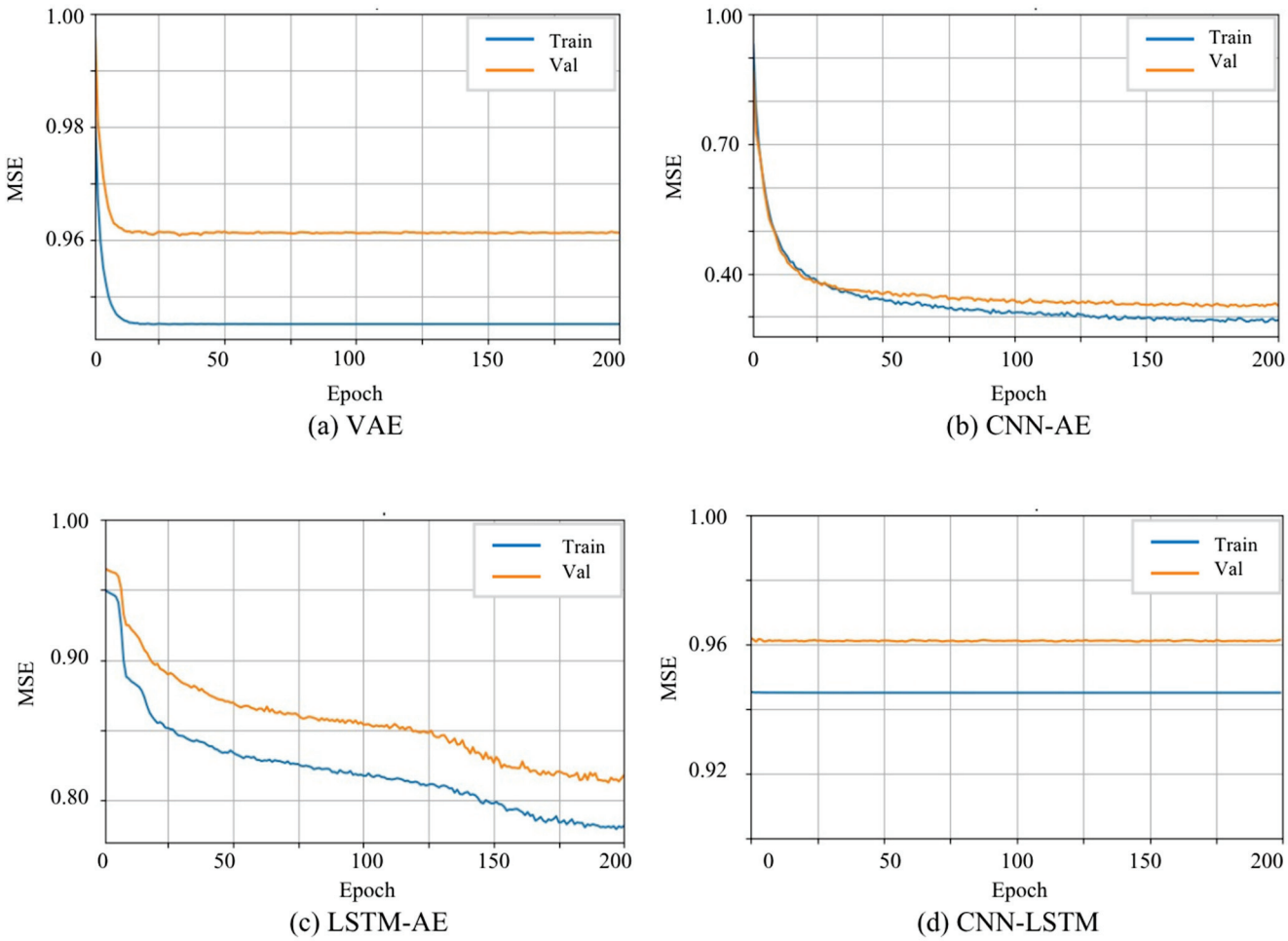


Figure 10. MSE over 200 epochs for the evaluated anomaly detection models.

The configured CNN–LSTM attempted to balance spatial and sequential processing, but the model exhibited the highest and most variable baseline error. While combustion events from wood and plastic were detectable, they lacked sufficient error separation for gasoline signals. While the CNN component successfully localized spatial channel perturbations, the LSTM module appears to have over-smoothed these signals during temporal modeling, causing transient anomalies to be suppressed. Figure 10 depicts training and validation losses across the full training cycle, highlighting generalization, error minimization, and overfitting risk for all models.

As shown in Figure 10, the CNN-AE exhibited rapid convergence, reducing training loss from approximately 0.90 to 0.30 and validation loss to 0.35 within the first 50 epochs. Both curves maintained a consistent, narrow gap, indicating rapid feature learning and minimal overfitting. The ability of the CNN-AE to extract local spatial and temporal features with minimal parameter overhead allowed for both fast convergence and high reconstruction accuracy, confirming its suitability for the task. In contrast, the VAE plateaued at a higher error (0.94 train/0.96 val), which is expected due to its prioritization of latent structure over raw reconstruction fidelity. This yielded smooth but less precise reconstructions, reducing sensitivity to short fire bursts. The LSTM-AE showed gradual convergence, with loss curves descending to ~0.80 (train) and ~0.85 (val). The widened gap is indicative of emerging overfitting to specific sequences. The CNN–LSTM exhibited suboptimal learning behavior, with both training and validation losses remaining nearly static throughout the optimization process, indicating that the model failed to meaningfully reduce reconstruction errors.

https://doi.org/10.3390/fire9050189

Model performance in detecting fire-related anomalies and suppressing false positives was evaluated using three indicators commonly used in unsupervised anomaly detection: False Alarm Rate (FAR), Overall Alarm Rate (OAR), and Average Run Length (ARL). FAR is the proportion of baseline (non-fire) windows that are incorrectly classified as anomalies, quantifying the frequency of false positives during normal system operation [70]. OAR denotes the percentage of fire windows correctly identified as anomalies [71]. This metric, which is equivalent to recall, reflects the detection sensitivity. ARL is the mean number of normal windows between two consecutive false alarms [72]. ARL serves as a reliability index and indicates how long the system can run without triggering false alerts. These metrics jointly measured how well each model detected fire-induced anomalies while avoiding false positives during normal conditions. The detection performance of the models is compared in Table 1.

As shown in Table 1, detection performance varied significantly across the four tested models when exposed to fire events from gasoline, wood, and plastic. CNN demonstrated perfect detection (100%) in all test cases, confirming its sensitivity to flame-induced signal disruptions. It also recorded the lowest FAR (3.6%), reflecting excellent stability under normal conditions. With an ARL of 24.3, the model maintained extended intervals between false positives. This performance was likely driven by the ability of the convolutional architecture to encode localized temporal–spectral structures with low reconstruction noise, creating sharp contrasts when anomalies occurred. LSTM-AE captured persistent anomalies (wood 100%, plastic 96.2%) but missed transients (gasoline 71.4%), revealing limited responsiveness to brief, low-intensity events. Its higher baseline jitter (FAR 4.9%, ARL 20.8) further indicated increased baseline noise and reduced stability. Nevertheless, its recurrent design allowed it to effectively capture prolonged anomalies through temporal modeling. In contrast, CNN-LSTM and VAE performed conservatively, excelling on wood and plastic, but underperforming on gasoline anomalies (28.6 and 14.3%, respectively).

Among all scenarios, gasoline fires presented the most challenging classification task due to their transient and subtle CSI distortions. Among all evaluated architectures, only CNN-AE reliably captured these anomalies. Its robust spatial filtering mechanisms yielded clear separation between normal and anomalous reconstructions, unlike the VAE and CNN–LSTM models, which failed to differentiate weak combustion signatures. In scenarios involving wood and plastic, the intensity and duration of CSI disturbances were sufficiently pronounced for all architectures to achieve successful detection. This is physically consistent with combustion behavior that hot gases, smoke aerosols, localized ionization, and turbulence modify the refractive properties of the propagation medium and disrupt the multipath field between the transmitter and receiver. This causes measurable CSI changes through attenuation, scattering, and path instability. The magnitude and duration of these effects vary with fuel type, producing distinct CSI signatures. In the present study, wood and plastic showed more persistent anomalies, likely due to greater smoke production, whereas gasoline caused weaker and shorter-lived channel disturbances.

Although these findings support the sensitivity of CSI to combustion-induced thermal and particulate effects, the individual physical contributions were not isolated in this study. Moreover, the experiments were performed under controlled line-of-sight conditions, with the fire source placed directly along the propagation path and without human activity, obstacles, or external RF interference. Although this configuration was suitable for validating the proposed concept, it represents an idealized environment and therefore limits direct extrapolation to practical residential settings. Future investigations should broaden the framework to include non-line-of-sight and multi-room configurations and incorporate human activity and environmental variability.

https://doi.org/10.3390/fire9050189

4. Conclusions

This study introduced a Wi-Fi-based early fire detection architecture that operated solely on CSI data, eliminating the need for thermal or gas sensors. By leveraging the ubiquity of wireless infrastructure, the framework reduced deployment complexity, expanded spatial coverage, and enabled early detection of combustion events. Exploiting ambient Wi-Fi signals, four unsupervised deep learning models were configured as fire detection systems to identify various combustion types based on perturbations in CSI signals induced by fire dynamics. The detection capabilities of these models were evaluated under three distinct combustion conditions: gasoline, wood, and plastic, each exhibiting different smoke and flame characteristics.

The CNN-AE was the most responsive architecture across all fire conditions. It achieved the best overall detection accuracy, maintained the lowest false alarm rate (3.6%), and posted the highest ARL among the evaluated models. Its 1D convolutional architecture effectively captured localized distortions in CSI signals with high precision. LSTM-AE was effective on slow-developing fires but less responsive to rapid anomalies. The VAE, while stable in normal operations, showed limited responsiveness to abrupt combustion signatures.

The CNN-AE model established a robust, low-latency framework for unsupervised fire detection using CSI from ambient Wi-Fi signals. Its generalization capacity enabled consistent performance across combustion sources with varying thermal and spectral dynamics. The architecture generalized well to diverse fuel types and demonstrated anomaly-resilient encoding properties, while its unsupervised formulation enhanced the scalability of RF-based anomaly systems. The results supported its application in residential safety systems and building automation frameworks.

This study was conducted under controlled conditions, focusing on autonomous fire detection in unoccupied environments. While this setup supports the intended use case, it limits applicability in more complex scenarios. Future work should extend the baseline dataset to include occupant motion and routine activities, thereby improving the precision of the system's ability to differentiate fire-related anomalies from other environmental disturbances. Additional directions include evaluation across diverse living environments and long-term deployment in real residential buildings to assess robustness under varying structural and interference conditions.

Author Contributions: Conceptualization, F.Y.O. and S.M.; Methodology, F.Y.O., S.M., A.C.A. and E.T.; Software, F.Y.O. and S.M.; Formal analysis, F.Y.O. and S.M.; Investigation, F.Y.O., S.M. and A.C.A.; Resources, F.Y.O., A.C.A. and E.T.; Data curation, F.Y.O. and S.M.; Writing—original draft, S.M.; Writing—review & editing, F.Y.O., S.M., A.C.A. and E.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original contributions presented in the study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: Author Ahmet Can Altunişik was employed by the company Dynamic Academy Software Construction Industry Trade Limited Company. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

https://doi.org/10.3390/fire9050189

References

1. Barua, U.; Mojtahedi, M.; Han, H.; Ansary, M.A.; Resilience, C. Sustainability and Smartness for Building Fire Risk Management. *Sustain. Dev.* **2025**, *33*, 1328–1354. [CrossRef]
2. Altın, M.; Kılıçarslan, M.F. Two real room fire and the evaluation of its consequences in terms of fire safety. *Case Stud. Therm. Eng.* **2023**, *44*, 102876. [CrossRef]
3. Rahman, F.S.; Tannous, W.K.; Avsar, G.; Agho, K.E.; Ghassempour, N.; Harvey, L.A. Economic Costs of Residential Fires: A Systematic Review. *Fire* **2023**, *6*, 399. [CrossRef]
4. Meacham, B.J. Fire safety of existing residential buildings: Building regulatory system gaps and needs. *Fire Saf. J.* **2023**, *140*, 103902. [CrossRef]
5. Godes, C.R.; Rodrigazo, S.A.; Cho, J.; Song, Y.; Yeon, J. Optimizing Evacuation Efficiency in Buildings: A BIM-Based Automated Approach to Sustainable Design. *Sustainability* **2024**, *16*, 9240. [CrossRef]
6. Jiang, L.; Shi, J.; Wang, C.; Pan, Z. Intelligent control of building fire protection system using digital twins and semantic web technologies. *Autom. Constr.* **2023**, *147*, 104728. [CrossRef]
7. Almatared, M.; Liu, H.; Abudayyeh, O.; Hakim, O.; Sulaiman, M. Digital-Twin-Based Fire Safety Management Framework for Smart Buildings. *Buildings* **2024**, *14*, 4. [CrossRef]
8. Xie, W.; Zeng, Y.; Zhang, X.; Wong, H.Y.; Zhang, T.; Wang, Z.; Wu, X.; Shi, J.; Huang, X.; Xiao, F.; et al. AIoT-powered building digital twin for smart firefighting and super real-time fire forecast. *Adv. Eng. Inform.* **2025**, *65*, 103117. [CrossRef]
9. Ehsan, I.; Mumtaz, A.; Khalid, M.I.; Iqbal, J.; Hussain, S.; Ullah, S.S.; Umar, F. Internet of Things-Based Fire Alarm Navigation System: A Fire-Rescue Department Perspective. *Mob. Inf. Syst.* **2022**, *2022*, 3830372. [CrossRef]
10. Ahn, Y.; Choi, H.; Kim, B.S. Development of early fire detection model for buildings using computer vision-based CCTV. *J. Build. Eng.* **2023**, *65*, 105647. [CrossRef]
11. Nazir, A.; Mosleh, H.; Takruri, M.; Jallad, A.-H.; Alhebsi, H. Early Fire Detection: A New Indoor Laboratory Dataset and Data Distribution Analysis. *Fire* **2022**, *5*, 11. [CrossRef]
12. Gunawaardena, A.E.; Ruwanthika, R.M.M.; Jayasekara, A.G.B.P. Computer vision based fire alarming system. In *Proceedings of the 2nd International Moratuwa Engineering Research Conference, MERCon*; IEEE: Piscataway, NJ, USA, 2016; Volume 2016, pp. 325–330. [CrossRef]
13. Tannous, W.K. The economic cost of unwanted automatic fire alarms. *Fire Saf. J.* **2021**, *124*, 103394. [CrossRef]
14. U.S. Fire Administration—FEMA. *Topical Fire Report Series Fatal Fires in Residential Buildings*; U.S. Fire Administration—FEMA: Washington, DC, USA, 2022.
15. Minegishi, Y.; Takahashi, W.; Ikehata, Y. Experimental study of occupants' situational awareness during a fire by sounding of wireless, interconnected alarm smoke detectors at a small-scale hostel that was originally a single-family house. *Jpn. Archit. Rev.* **2024**, *7*, e12453. [CrossRef]
16. Kincaid, S. Fire prevention in historic buildings—approaches for safe practice. *Hist. Environ. Policy Pract.* **2022**, *13*, 361–380. [CrossRef]
17. Wang, L.; Zhang, X.; Li, L.; Li, B.; Mei, Z. Experimental Study on Early Fire Smoke Characteristics in a High-Volume Space: A Fire Detection Perspective. *Fire* **2024**, *7*, 298. [CrossRef]
18. Yuan, F. Video-based smoke detection with histogram sequence of LBP and LBPV pyramids. *Fire Saf. J.* **2011**, *46*, 132–139. [CrossRef]
19. Geetha, S.; Abhishek, C.S.; Akshayanat, C.S. Machine Vision Based Fire Detection Techniques: A Survey. *Fire Technol.* **2021**, *57*, 591–623. [CrossRef]
20. Muhammad, K.; Ahmad, J.; Mehmood, I.; Rho, S.; Baik, S.W. Convolutional Neural Networks Based Fire Detection in Surveillance Videos. *IEEE Access* **2018**, *6*, 18174–18183. [CrossRef]
21. Saponara, S.; Elhanashi, A.; Gagliardi, A. Real-time video fire/smoke detection based on CNN in antifire surveillance systems. *J. Real Time Image Process.* **2021**, *18*, 889–900. [CrossRef]
22. Gragnaniello, D.; Greco, A.; Sansone, C.; Vento, B. FLAME: Fire detection in videos combining a deep neural network with a model-based motion analysis. *Neural Comput. Appl.* **2025**, *37*, 6181–6197. [CrossRef]
23. Bu, F.; Gharajeh, M.S. Intelligent and vision-based fire detection systems: A survey. *Image Vis. Comput.* **2019**, *91*, 103803. [CrossRef]
24. Yar, H.; Khan, Z.A.; Rida, I.; Ullah, W.; Kim, M.J.; Baik, S.W. An efficient deep learning architecture for effective fire detection in smart surveillance. *Image Vis. Comput.* **2024**, *145*, 104989. [CrossRef]
25. Yang, X.; Li, Y.; Chen, Q. Automated image-based fire detection and alarm system using edge computing and cloud-based platform. *Internet Things* **2024**, *28*, 101402. [CrossRef]
26. Pincott, J.; Tien, P.W.; Wei, S.; Calautit, J.K. Indoor fire detection utilizing computer vision-based strategies. *J. Build. Eng.* **2022**, *61*, 105154. [CrossRef]

https://doi.org/10.3390/fire9050189

27. Mondal, M.S.; Prasad, V.; Kumar, R.; Saha, N.; Guha, S.; Ghosh, R.; Mukhopadhyay, A.; Sarkar, S. Automating Fire Detection and Suppression with Computer Vision: A Multi-Layered Filtering Approach to Enhanced Fire Safety and Rapid Response. *Fire Technol.* **2023**, *59*, 1555–1583. [CrossRef]
28. Piera, J.Y.; Salva, J.K.G. A Wireless Sensor Network for Fire Detection and Alarm System. In *Proceedings of the 2019 7th International Conference on Information and Communication Technology (ICoICT)*; IEEE: Piscataway, NJ, USA, 2019; pp. 1–5.
29. Yoddumnern, A.; Yooyativong, T.; Chaisricharoen, R. The wifi multi-sensor network for fire detection mechanism using fuzzy logic with IFTTT process based on cloud. In Proceedings of the ECTI-CON 2017—2017 14th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, Phuket, Thailand, 27–30 June 2017; pp. 785–789. [CrossRef]
30. Shah, R.; Satam, P.; Sayyed, M.A.; Salvi, P. Wireless Smoke Detector and Fire Alarm System. *Int. Res. J. Eng. Technol.* **2019**, *6*, 1407–1412. Available online: https://www.irjet.net/archives/V6/i1/IRJET-V6I1258.pdf (accessed on 24 October 2025).
31. Ming, X.-C. Smart Fire Rescue System Design Based on WI-FI Mode. *DEStech Trans. Comput. Sci. Eng.* **2017**. [CrossRef]
32. Lakshmi, G.; Akash, T.; Gowtham, J.; Venkatesh, R.; Abhinesh, S.B. Revolutionizing Home Fire Safety: IoT-Based Wireless Fire Alarm System for Easy Installation and Rapid Response. In *Proceedings of the 2023 Intelligent Computing and Control for Engineering and Business Systems, ICCEBS 2023*; IEEE: Piscataway, NJ, USA, 2023. [CrossRef]
33. Wu, D.; Zeng, Y.; Gao, R.; Li, S.; Li, Y.; Shah, R.C.; Lu, H.; Zhang, D. WiTraj: Robust Indoor Motion Tracking with WiFi Signals. *IEEE Trans. Mob. Comput.* **2023**, *22*, 3062–3078. [CrossRef]
34. Zhang, Y.; Wang, X.; Wen, J.; Zhu, X. WiFi-based non-contact human presence detection technology. *Sci. Rep.* **2024**, *14*, 3605. [CrossRef]
35. Guo, Z.; Yuan, W.; Gui, L.; Sheng, B.; Xiao, F. BreatheBand: A Fine-grained and Robust Respiration Monitor System Using WiFi Signals. *ACM Trans. Sens. Netw.* **2023**, *19*, 3582079. [CrossRef]
36. Bao, N.; Du, J.; Wu, C.; Hong, D.; Chen, J.; Nowak, R.; Lv, Z. Wi-Breath: A WiFi-Based Contactless and Real-Time Respiration Monitoring Scheme for Remote Healthcare. *IEEE J. Biomed. Health Inform.* **2023**, *27*, 2276–2285. [CrossRef] [PubMed]
37. Naticchia, B.; Vaccarini, M.; Carbonari, A. A monitoring system for real-time interference control on large construction sites. *Autom. Constr.* **2013**, *29*, 148–160. [CrossRef]
38. Guo, R.; Li, H.; Han, D.; Liu, R. Feasibility Analysis of Using Channel State Information (CSI) Acquired from Wi-Fi Routers for Construction Worker Fall Detection. *Int. J. Environ. Res. Public Health* **2023**, *20*, 4998. [CrossRef]
39. Stojanović, G.; Radovanović, M.; Malešev, M.; Radonjanin, V. Monitoring of Water Content in Building Materials Using a Wireless Passive Sensor. *Sensors* **2010**, *10*, 4270–4280. [CrossRef]
40. Gacem, M.A.; Zakaria, A.S.; Ismail, M.H.; Tariq, U.; Yehia, S. Measurement of Construction Materials Properties Using Wi-Fi and Convolutional Neural Networks. *IEEE Access* **2022**, *10*, 126100–126116. [CrossRef]
41. Morgenthal, G.; Eick, J.F.; Rau, S.; Taraben, J. Wireless Sensor Networks Composed of Standard Microcomputers and Smartphones for Applications in Structural Health Monitoring. *Sensors* **2019**, *19*, 2070. [CrossRef]
42. Toledo, E.; Cury, A.; Landre, J. Assessment of low-cost wireless sensors for structural health monitoring applications. *Rev. IBRACON Estrut. Mater.* **2021**, *14*, e14213. [CrossRef]
43. Zhong, S.; Huang, Y.; Ruby, R.; Wang, L.; Qiu, Y.X.; Wu, K. Wi-fire: Device-free fire detection using WiFi networks. In *Proceedings of the IEEE International Conference on Communications*; IEEE: Piscataway, NJ, USA, 2017. [CrossRef]
44. Li, J.; Sharma, A.; Mishra, D.; Seneviratne, A. Fire Detection Using Commodity WiFi Devices. In *Proceedings of the 2021 IEEE Global Communications Conference (GLOBECOM)*; IEEE: Piscataway, NJ, USA, 2022; pp. 1–6. [CrossRef]
45. Sharp, I.; Yu, K. Indoor WiFi Positioning. In *Wireless Positioning: Principles and Practice*; Navigation: Science and Technology; Springer: Singapore, 2019; pp. 219–240. [CrossRef]
46. Wang, W.H.; Wang, B.; Hu, Y.; Zhu, G.; Liu, K.J.R. Passive WiFi-Based Indoor Localization Through Multi-Domain Features. In *Proceedings of the 2024 IEEE 10th World Forum on Internet of Things, WF-IoT*; IEEE: Piscataway, NJ, USA, 2024; pp. 530–535. [CrossRef]
47. Li, J.; Sharma, A.; Mishra, D.; Seneviratne, A. Thermal Profiling by WiFi Sensing in IoT Networks. In *Proceedings of the IEEE Global Communications Conference, GLOBECOM*; IEEE: Piscataway, NJ, USA, 2021. [CrossRef]
48. Li, J.; Yingchanakiate, K.; Mishra, D.; Seneviratne, A. Thermal Source Localization Using WiFi Sensing. In *Proceedings of the 2024 IEEE Global Communications Conference (GLOBECOM 2024)*; IEEE: Piscataway, NJ, USA, 2024; pp. 1875–1880. [CrossRef]
49. Yuvaraj, R.; Bhalerao, S.A.; Murugesan, K.; Vellaiyan, S.; Van Minh, N. Real-time fire detection and suppression system using YOLO11n and Raspberry Pi for thermal safety applications. *Case Stud. Therm. Eng.* **2025**, *75*, 107159. [CrossRef]
50. Ahmed, H.F.T.; Ahmad, H.; Phang, S.K.; Harkat, H.; Narasingamurthi, K. Wi-Fi CSI Based Human Sign Language Recognition using LSTM Network. In *Proceedings of the 2021 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology, IAICT*; IEEE: Piscataway, NJ, USA, 2021; pp. 51–57. [CrossRef]

https://doi.org/10.3390/fire9050189

51. Lin, S.; Clark, R.; Birke, R.; Schonborn, S.; Trigoni, N.; Roberts, S. Anomaly Detection for Time Series Using VAE-LSTM Hybrid Model, ICASSP. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Barcelona, Spain, 4–8 May 2020; pp. 4322–4326. [CrossRef]
52. Sruthi, P.; Udgata, S.K. Wi-Fi sensing based person identification and activity recognition using two-phase deep learning model. *Eng. Appl. Artif. Intell.* **2024**, *132*, 107904. [CrossRef]
53. Hoang, M.T.; Yuen, B.; Ren, K.; Dong, X.; Lu, T.; Westendorp, R.; Reddy, K. A CNN-LSTM Quantifier for Single Access Point CSI Indoor Localization. *arXiv* **2020**, arXiv:2005.06394. [CrossRef]
54. Pereira, J.; Silveira, M. Unsupervised Anomaly Detection in Energy Time Series Data Using Variational Recurrent Autoencoders with Attention. In *Proceedings of the 17th IEEE International Conference on Machine Learning and Applications, ICMLA*; IEEE: Piscataway, NJ, USA, 2018; pp. 1275–1282. [CrossRef]
55. Aparna, G.B.; Poulose, A. Human Activity Recognition via Wi-Fi CSI and VAE-CNN-BiLSTM Hybrid Deep Learning. In *Proceedings of 2024 International Conference on Brain Computer Interface and Healthcare Technologies, ICon-BCIHT*; IEEE: Piscataway, NJ, USA, 2024; pp. 64–69. [CrossRef]
56. Malhotra, P.; Ramakrishnan, A.; Anand, G.; Vig, L.; Agarwal, P.; Shroff, G. LSTM-based Encoder-Decoder for Multi-sensor Anomaly Detection. *arXiv* **2016**, arXiv:1607.00148.
57. Khanmohammadi, F.; Azmi, R. Time-Series Anomaly Detection in Automated Vehicles Using D-CNN-LSTM Autoencoder. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 9296–9307. [CrossRef]
58. Dewor, R.; Schulz, C.; Büttgen, R.D.; Brands, T.; Heufer, K.A.; Koß, H.J. Development of the first Raman scattering thermometry during the first stage ignition in a rapid compression machine and determination of detection limits for NO-LIF. *Appl. Energy Combust. Sci.* **2023**, *16*, 100228. [CrossRef]
59. Üstek, İ.; Arana-Catania, M.; Farr, A.; Petrunin, I. Deep Autoencoders for Unsupervised Anomaly Detection in Wildfire Prediction. *Earth Space Sci.* **2024**, *11*, e2024EA003997. [CrossRef]
60. Rassam, M.A. Autoencoder-Based Neural Network Model for Anomaly Detection in Wireless Body Area Networks. *IoT* **2024**, *5*, 852–870. [CrossRef]
61. Introduction to ESP-CSI Solution—Technical Documents, Shanghai, 2016. Available online: https://docs.espressif.com/projects/esp-techpedia/en/latest/ESP-Techpedia-en-master.pdf (accessed on 10 September 2025).
62. Hutar, M.; Brida, P.; MacHaj, J. Performance of MLP NN Using CSI and Different Standard IEEE 802.11 with Variable Channel Width. In Proceedings of the 34th International Conference Radioelektronika, Zilina, Slovakia, 17–18 April 2024. [CrossRef]
63. Dissanayake, C.M.; Halgamuge, M.N.; Ramamohanarao, K.; Moran, B.; Farrell, P. The signal propagation effects on IEEE 802.15.4 radio link in fire environment. In *Proceedings of the 2010 Fifth International Conference on Information and Automation for Sustainability*; IEEE: Piscataway, NJ, USA, 2010; pp. 411–414. [CrossRef]
64. Dang, X.; Si, X.; Hao, Z.; Huang, Y. A Novel Passive Indoor Localization Method by Fusion CSI Amplitude and Phase Information. *Sensors* **2019**, *19*, 875. [CrossRef]
65. Zhang, D.; Yao, L.; Chen, K.; Huang, C.; Awad, H.; Ali, H.; Seytnazarov, S. Human Walking Direction Detection Using Wireless Signals, Machine and Deep Learning Algorithms. *Sensors* **2023**, *23*, 9726. [CrossRef]
66. Schäfer, J.; Barrsiwal, B.R.; Kokhkharova, M.; Adil, H.; Liebehenschel, J. Human Activity Recognition Using CSI Information with Nexmon. *Appl. Sci.* **2021**, *11*, 8860. [CrossRef]
67. Pearson, R.K.; Neuvo, Y.; Astola, J.; Gabbouj, M. Generalized Hampel Filters. *EURASIP J. Adv. Signal Process.* **2016**, *2016*, 87. [CrossRef]
68. Wu, K.; Zhong, S.; Wang, L. Indoor Fire Detection and Alarm Method and System Based on Wireless Network Signal Transmission. Patent WO2017117910A1, 13 July 2017.
69. Okur, F.Y.; Altunişik, A.C.; Okur, E.K. A Novel Approach for Anomaly Detection in Vibration-Based Structural Health Monitoring Using Autoencoders in Deep Learning. *Struct. Control Health Monit.* **2025**, *2025*, 5602604. [CrossRef]
70. Reehorst, E.T.; Schniter, P. Score Combining for Contrastive OOD Detection. *arXiv* **2025**, arXiv:2501.12204. [CrossRef]
71. Laxhamma, R. Anomaly Detection in Trajectory Data for Surveillance Applications. Ph.D. Thesis, Örebro University, Örebro, Sweden, 2011.
72. Fisch, A.T.M.; Bardwell, L.; Eckley, I.A. Real time anomaly detection and categorisation. *Stat. Comput.* **2022**, *32*, 55. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

https://doi.org/10.3390/fire9050189

remote sensing

MDPI

Article

Geographic Bias Analysis and Cross-Domain Generalization in Deep Learning-Based Building Damage Assessment

Shruti Kshirsagar [1], Bharath Chandra [1], Unaza Tallal [1,*], Rajiv Bagai [1] and Atri Dutta [2]

[1] School of Computing, Wichita State University, Wichita, KS 67260, USA; shruti.kshirsagar@wichita.edu (S.K.); bxparupati@shockers.wichita.edu (B.C.); rajiv.bagai@wichita.edu (R.B.)
[2] Department of Aerospace Engineering, Wichita State University, Wichita, KS 67260, USA; atri.dutta@wichita.edu
* Correspondence: uxtallal@shockers.wichita.edu

Highlights

What are the main findings?

- AI-based building damage assessment models have weak generalization across geographically diverse disaster locations, and their performance is strongly affected by geographic and structural differences.
- Structural feature enhancement using auxiliary channels is critical for improving robustness, and the proposed methods increased the overall F1 score, with especially large gains in the minor and major damage classes.

What are the implications of the main findings?

- Simply increasing the amount of training data is not enough; model design must account for environmental context, architectural diversity, and terrain complexity to reduce geographic bias.
- Domain adaptation methods such as Supervised fine-tuning and Unsupervised CORAL can substantially improve performance in unseen locations, making AI damage assessment more reliable for real-world disaster response.

Abstract

Automated building damage assessment from satellite imagery has become increasingly critical for rapid disaster response and humanitarian relief operations. However, current state-of-the-art deep learning models exhibit significant generalization challenges when deployed to geographically and environmentally diverse regions. This study investigates the nature and extent of geographic bias in building damage detection systems, revealing that model performance degradation occurs primarily from geographic and structural characteristics rather than insufficient training data representation. Through a systematic evaluation of top-performing xView2 competition solutions across 17 disaster locations across multiple climate zones, we found that even state-of-the-art models struggle with generalization under geographic shift, particularly for the minor and major damage classes, and exhibit strong geographic biases toward certain regions. In this work, we explore a six-channel Fusion Augmentation strategy that enriches RGB imagery with auxiliary structural enhancement channels, together with supervised fine-tuning and unsupervised CORAL-based domain adaptation for three unseen regions. The experimental results demonstrate a substantial improvement of 7.1% overall F1 score, with notable gains for intermediate damage categories such as minor and major damage. Domain adaptation experiments on three unseen locations show that combining Fusion Augmentation with supervised fine-tuning yields 40.8% and 60.0% improvements over the minor and major classes, while unsupervised CORAL achieves 24.2% and 39.5% improvements over the

minor and major damage classes compared to benchmarks. These findings highlight persistent geographic bias and demonstrate that structural feature enhancement combined with domain adaptation is essential for globally deployable damage assessment systems.

Keywords: building damage assessment; satellite imagery; artificial intelligence; deep learning; geographic bias; domain adaptation; data augmentation; cross-domain generalization; disaster response; remote sensing; transfer learning; computer vision

1. Introduction

Climate change is increasing the frequency and intensity of natural disasters such as hurricanes, earthquakes, floods, and wildfires, affecting communities worldwide [1]. Natural disasters globally displaced 26.4 million people in 2023, causing economic damages of roughly USD 380 billion [2]. Emergency action following such events is crucial, as a timely and precise assessment of building damage can save trapped survivors. Due to infrastructure damage, safety hazards, and the scale of affected areas, manual field inspections by structural engineers are generally insufficient for damage assessment. These constraints necessitate automated, scalable, and reliable post-disaster building damage assessment. Satellite and airborne photos provide speedy, wide-area coverage of disaster-affected areas without endangering assessment crews [3]. Within hours of a tragedy, modern Earth observation satellites may acquire high-resolution photos of structural damage across cities or regions. Manual interpretation of this footage is challenging and time-consuming, especially when studying hundreds of thousands of buildings in large catastrophe zones. This bottleneck slows emergency response decision making, which is crucial for search-and-rescue and resource allocation.

Recent advancements in deep learning (DL), especially CNNs, have enabled automated detection of building damage from satellite images [4]. These models can analyze large volumes of visual data and classify damage severity with accuracy that approaches or exceeds that of human specialists. The xView2 Challenge, organized by the Defense Innovation Unit (DIU), has advanced building damage assessment by providing the largest publicly available dataset of annotated satellite imagery, spanning 19 disasters and 850,000 labeled buildings [5]. Top-performing solutions from this competition achieved F1 scores above 0.85 for building location and damage classification, proving automated damage assessment systems are technically feasible. Despite these impressive benchmark results, artificial intelligence systems for building damage assessment face critical challenges when deployed across geographically and environmentally diverse regions. Models trained on specific disasters often fail to generalize due to domain shift, systematic differences in data distributions caused by varying architectural styles, environmental conditions, and disaster types. Recent operational deployments by humanitarian organizations have revealed that models achieving 90% accuracy in North American contexts may perform at only 30–40% accuracy in South Asian or tropical environments, highlighting fundamental limitations in AI generalization.

Despite impressive benchmark results, a critical gap remains between controlled evaluation and real-world deployment: deep learning models trained on specific disasters frequently fail when applied to new geographic regions, different disaster types, or unfamiliar architectural contexts [5–7]. This degradation arises from geographic skew in training data, class imbalance favoring undamaged structures, and climate zone variability, not merely from insufficient data volume [5]. The consequences are severe: models achieving high accuracy on North American hurricanes have shown dramatic performance drops

https://doi.org/10.3390/rs18101529

when deployed to earthquake-affected regions in South Asia, as documented in operational use by the United Nations and World Bank [8]. A system that systematically underestimates damage in tropical or Mediterranean environments risks delayed humanitarian aid and preventable loss of life.

Our work addresses this gap by examining geographic bias in AI-based building damage assessment and investigating whether structural feature enhancement and domain adaptation can improve cross-location robustness. Specifically, we investigate these challenges through systematic empirical analysis across locations and assess whether structurally enriched input representations, combined with domain adaptation, can improve transfer performance under geographic shifts. Our main contributions are as follows:

1. Geographic bias analysis: We conduct a systematic evaluation of the top-performing xView2 solutions across 17 disaster locations and show substantial variation in performance across regions and damage categories. The results indicate that cross-location generalization cannot be explained by training sample quantity alone and is strongly affected by geographic, structural, and environmental differences.
2. Integrated augmentation–adaptation framework: We systematically evaluate supervised fine-tuning and unsupervised CORAL domain adaptation combined with Fusion Augmentation across three out-of-domain locations (Texas, Ayiti, Portugal), while also examining whether structurally enriched input representations improve transfer performance under geographic domain shift.

The remainder of this paper is organized as follows: Section 2 reviews related work in building damage classification, data augmentation, domain adaptation, and model bias. Section 3 discusses methods and material, including the dataset, distribution, our approach, Fusion Augmentation, and domain adaptation methodologies, experimental setup, and evaluation protocols. Section 4 presents in-domain and out-of-domain test results and geographic bias patterns. Section 5 concludes with an overview of contributions and their potential impact on operational disaster response systems.

2. Related Work

This section reviews recent advancements in deep learning models for building damage detection, focusing on damage classification approaches, model biases, data augmentation strategies, and domain adaptation techniques.

2.1. State-of-the-Art Models for Building Damage Detection

Deep learning and high-resolution Earth observation data have advanced satellite-based building damage classification. Satellite imagery is now widely used by the International Charter and Copernicus Emergency Management Service for disaster damage evaluation and reference maps [9]. The use of convolutional neural networks (CNNs) for object segmentation [10] has led to robust capabilities in automated building damage detection [11]. Two of the main types of method for assessing building damage are bi-temporal methods (analyzing pre- and post-disaster imagery) [11,12] and single-image methods (assessing damage using only post-disaster imagery) [13,14]. However, bi-temporal techniques require temporally aligned image pairs to measure changes and give more thorough damage assessment information. Single-image methods are faster and useful for sudden, unexpected disasters when pre-disaster imagery is unavailable. Following the publication of the xBD dataset and the xView2 Challenge, building damage detection models have quickly evolved to focus on architectural improvements, class-imbalance handling, and cross-domain generalization. Kaur et al. [15] developed DAHiTrA, a transformer-based approach that maps pre- and post-disaster images into a common feature space using hierarchical spatial features, capturing long-range dependencies crucial for damage assessment.

https://doi.org/10.3390/rs18101529

The Ordinal Class Distance Penalty Loss (OCDPL) by Tsai and Lin [16] addresses class imbalance by penalizing misclassifications based on the difference between predicted and actual damage levels. This approach uses the damage category structure to improve minority-class performance. According to Wiguna et al. [7], a semi-supervised framework that uses pseudo-labeling and iterative fine-tuning improved accuracy by 21% on the Noto Peninsula Earthquake dataset, highlighting the potential of leveraging abundant unlabeled disaster data. Jakubik et al. [17] developed foundation models for Earth monitoring, using large-scale pretraining and unsupervised domain adaptation to produce transferable representations that adapt to new locations and hazards with minimal fine-tuning. More recent work has continued this trend. Ref. [18] proposed DAVI, a vision foundation model-based framework for building damage assessment that does not requiring target-region labels on unseen regions. Ref. [19] introduced ViPDE; the study uses the Segment Anything Model (SAM) and visual prompt learning for post-disaster damage evaluation. Also, Ref. [20] proposed GAMSF, a transformer-based spatial–frequency fusion network, and validated its transferability on an independent benchmark.

2.2. Bias and Generalization Challenges

Despite impressive performance on benchmark datasets, deep learning models for building damage detection face significant challenges related to bias and limited generalization. Bias refers to systematic errors that cause models to perform inconsistently across different datasets or deployment scenarios, particularly when the training data fails to represent the full variety of real-world conditions [21]. Many types of bias have been recognized [22]: Class imbalance bias is caused by the xBD dataset's imbalance, where undamaged buildings outnumber damaged ones, leading to model bias towards the majority class [5,7,11,15]. Geographic bias arises when training data is concentrated in North America, leading models to encode region-specific building materials, architectural styles, and landscape attributes. Melamed et al. [21] found that damaged structures are spatially grouped, with building damage levels substantially associated with surrounding buildings. The top-5 xView2 solutions perform well in highly damaged locations but poorly in isolated damaged structures, which could lead to important oversights and delayed emergency assistance. Real-world applicability is further limited by sensor- and image-quality biases arising from satellite sensors, atmospheric conditions, and acquisition parameters. These biases drive our research on data augmentation and domain adaptation to improve generalization.

2.3. Data Augmentation Techniques

Data augmentation is essential for deep learning, especially in resource-constrained domains such as building damage detection. Research from various fields shows that effective augmentation must be task-specific, preserve crucial traits, and offer meaningful variability. Augmentation in computer vision relies on traditional geometric transformations, such as flips, rotations, scale adjustments, and color jittering [23,24]. Adedeji et al. [23] systematically evaluated these techniques for satellite image classification, noting that while they improve robustness to orientation and perspective changes, their impact is somewhat limited for satellite imagery due to inherent uniformity. DCGAN-based generative techniques for satellite image synthesis are more effective at producing diverse instances but require more computational resources [23]. Advanced techniques include fusion-based augmentation, where auxiliary information channels combine with RGB data to enrich representations [25]; RICAP, which patches multiple images together [26]; the Albumentations library, offering efficient transformations [27]; and Cutout, which masks regions to force broader context utilization [28]. Critically, research demonstrates that effective

https://doi.org/10.3390/rs18101529

augmentation must be quality-aware and task-specific: strategies should adapt to input quality characteristics [29], work synergistically with representation learning [30], and most importantly, preserve task-relevant features while introducing meaningful variability [31]. For building damage assessment, this implies retaining structural boundaries while enhancing environmental resilience. Advances in building damage identification have not addressed structural edge preservation and enhancement, which are vital for distinguishing damage levels, especially in "minor" and "major" categories [32,33].

2.4. Domain Adaptation Methods

Domain adaptation transfers knowledge from source domains (training data) to target domains (deployment scenarios) with different data distributions. Supervised approaches utilize labeled target samples for fine-tuning, while unsupervised methods align representations without labels [34–36]. Jakubik et al. [17] used unsupervised domain adaptation, combining pretraining on related tasks with zero-shot adaptation to new hazard types, thereby enhancing generalizability and reducing the need for labeled data. UniDA allows adaptation without target label set knowledge [34], while source-free adaptation works when source data is unavailable [35]. Successful adaptation requires both feature distribution alignment and preservation of task-relevant characteristics, with effectiveness enhanced when combined with task-specific augmentation under significant domain shift [29,30,36]. Hu and Tang [35] found that employing CycleGAN and adversarial training for transfer learning and domain adaptation enhanced cross-disaster generalization for building damage assessment across 14 events. Recent research by Parupati et al. [32,33] suggests that combining augmentation and adaptation leads to better generalization than either strategy alone. In line with this, Mouradi et al. [37] investigated cross-disaster building damage detection using a two-stage ensemble framework with supervised domain adaptation. The results showed that supervised domain adaptation was essential for robust four-class damage classification in unseen geographic regions.

3. Materials and Methods

This section describes the dataset, preprocessing procedures, fusion augmentation approach, domain adaptation techniques, baseline models, and evaluation metrics.

Figure 1 presents the overall pipeline of the proposed framework for building damage detection and geographic bias analysis. First, the source-domain pre- and post-disaster images are preprocessed. Next, Fusion Augmentation is applied to enrich the original input representation with auxiliary structural channels. The resulting enhanced representation is then used to train the source model. Finally, the framework supports two downstream analyses: in-domain geographic bias analysis across 17 locations and cross-domain adaptation to unseen target regions using supervised or unsupervised adaptation methods.

3.1. Dataset

For building damage assessment, the xBD (xView2 Building Damage) dataset [5] is the largest publicly available dataset from satellite imagery. The dataset consists of paired pre- and post-disaster images collected by Maxar/DigitalGlobe satellites (Maxar Technologies Inc., Westminster, CO, USA) at a spatial resolution of 0.3–0.8 m. The dataset covers approximately 45,362 square kilometers and includes more than 850,000 annotated building polygons, collected from 19 disaster events across five continents. A wide range of hazard types are included in the xBD dataset, such as severe wildfires (Santa Rosa, Woolsey, Pinery, SoCal, Portugal), major hurricanes (Michael, Harvey, Matthew, Florence), destructive tornadoes (Moore, Joplin), volcanic eruptions (Guatemala, Lower Puna), earthquakes (Nepal, Mexico), floods (Midwest), and the Palu tsunami. These events take place across

https://doi.org/10.3390/rs18101529

diverse climate zones, including tropical, volcanic, arid, temperate, and mountainous regions, making xBD one of the most diverse datasets available. Each building instance is labeled with a polygon footprint, and a damage level is assigned according to the four-tier Joint Damage Scale, which includes four categories: No damage, minor damage, major damage, and destroyed. The dataset also offers global disaster maps of affected locations and illustrative pre- and post-disaster image pairs.

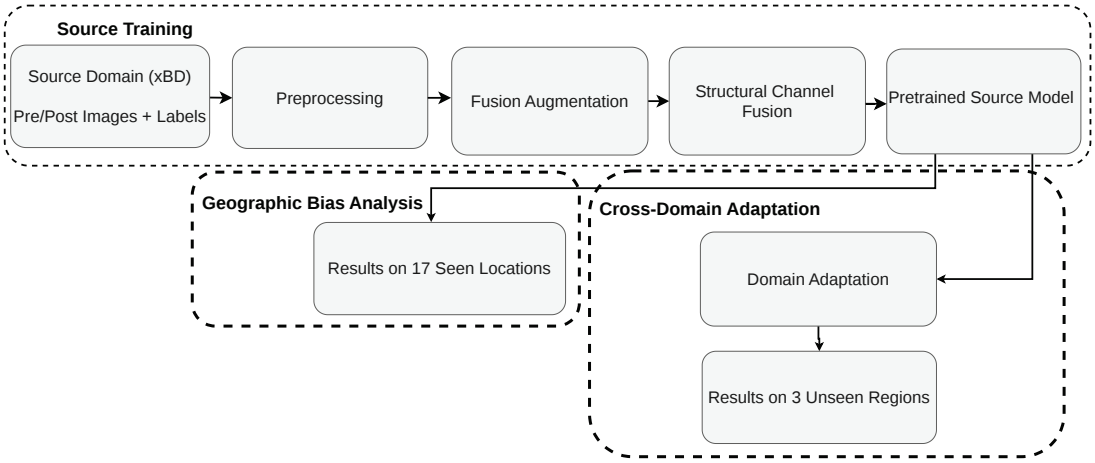


Figure 1. End-to-end pipeline of the proposed framework for geographic bias analysis and cross-domain building damage assessment.

Table 1 outlines four damage classes presented in the xBD dataset, ranging from undamaged buildings to completely destroyed structures. These classes describe the level of damage to each building, ranging from no structural issues to full collapse or destruction. The xBD dataset has a severe damage-class distribution imbalance, a pattern commonly seen in damage-related datasets. Table 2 presents the frequency of each category across the 19 disaster events. Out of a total of 411,357 labeled buildings, 76% fall in the no damage class, whereas minor damage, major damage, and destroyed make up only 7–9% of the whole dataset. This skewed distribution underscores the class imbalance challenge when training models on the xBD dataset.

Table 1. Damage scale descriptions. Adapted from the xBD dataset [5], used under the Creative Commons license stated by the dataset source and the table formatting was prepared by the authors.

Damage Class	Description
No damage	Undisturbed. No sign of water, structural or single damage, or burn marks.
Minor damage	Building partially burnt, water surrounding the structure, roof elements missing, or visible cracks.
Major damage	Partial wall or roof collapse, encroaching, surrounded by mud/water.
Destroyed	Completely collapsed, partially or completely covered with water/mud, or no longer present.

https://doi.org/10.3390/rs18101529

Table 2. Class distribution in xBD dataset.

Damage Class	Count	Percentage
No damage	313,033	76.2%
Minor damage	36,860	9.0%
Major damage	29,904	7.3%
Destroyed	31,560	7.5%
Total	411,357	100%

Furthermore, the distribution of samples varies significantly across disaster types and geographic regions, as illustrated in Figure 2. Some locations, such as Uttar Pradesh (flooding), Indonesia (tsunami), Mexico (earthquake), and California (fire), contribute a comparatively large number of labeled buildings, whereas others, including Hawaii (volcano), Guatemala (volcano), and South Australia (fire), are represented by substantially fewer samples. This geographic skew introduces an inherent bias, which is the central focus of our investigation.

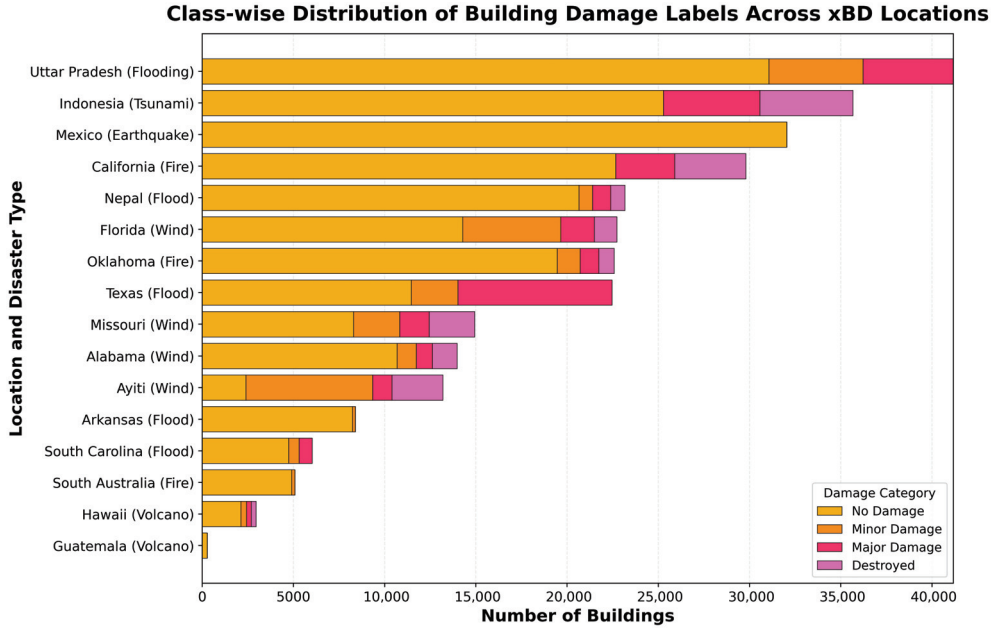


Figure 2. Class-wisedistribution of building damage labels across xBD locations used in this study. The figure highlights class imbalance and regional variation in sample composition [5].

3.2. Image Preprocessing

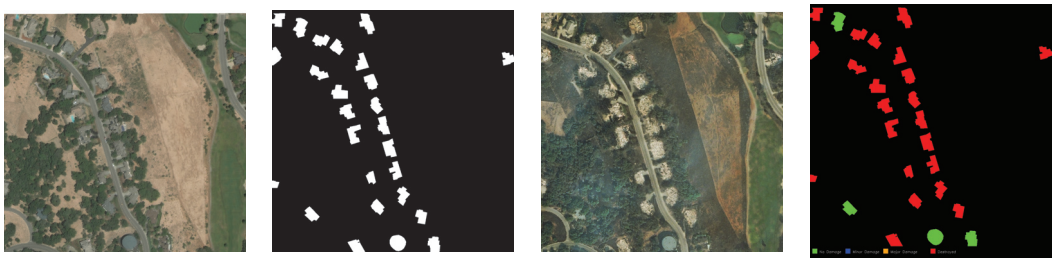
The xBD dataset includes building damage annotations in GeoJSON format, each containing polygon coordinates for a structure, along with metadata such as damage level, disaster type, and image identifiers. We converted these annotations to pixel-wise masks for model training using the following pipeline.

3.2.1. Polygon Extraction and Mask Generation

First, we parsed the GeoJSON files to extract the polygon coordinates (x,y), where each point represents a vertex of a building. We then converted these geographic coordinates into pixel coordinates using the geotransformation information of each image. Next, we generated masks for each building polygon using OpenCV's fillPoly function. For building localization, pixels belonging to buildings were labeled as 1, and background pixels as 0. For damage classification, labels 0,1,2,3,4 represent background, no damage, minor, major,

https://doi.org/10.3390/rs18101529

and destroyed, respectively. When multiple buildings overlapped at the same pixel, we assigned the label with the highest damage level. This ensured that regions with severe damage were properly captured. Figure 3 shows the full preprocessing pipeline from raw GeoJSON data to final masks.



(**a**) (**b**) (**c**) (**d**)

Figure 3. Pre-disaster and post-disaster images of the Santa Rosa wildfire with their corresponding masks. The pre- and post-disaster satellite images are from the publicly available xBD dataset [5] used under the Creative Commons Attribution-Non Commercial-ShareAlike 4.0 International license. The corresponding masks were generated by the authors. (**a**) Pre-disaster image. (**b**) Pre-disaster mask. (**c**) Post-disaster image. (**d**) Post-disaster mask.

3.2.2. Image Normalization

All satellite imagery is normalized using ImageNet statistics [38] to enable transfer learning from pretrained encoders.

$$I_{\text{norm}} = \frac{I - \mu}{\sigma} \tag{1}$$

where $\mu = [0.485, 0.456, 0.406]$ and $\sigma = [0.229, 0.224, 0.225]$ represent the mean and standard deviation for the RGB channels. All images are downsized to 1024×1024 pixels, with zero-padding for smaller ones. To process images larger than 1024×1024, we adopt a sliding-window technique with 50% overlap and aggregate predictions via majority voting during inference.

3.3. Fusion Augmentation

Post-disaster satellite imagery presents significant challenges for AI-based damage detection: structural edges are obscured by debris, contrast is reduced by atmospheric or illumination effects, fine-grained damage features are masked by environmental factors, and buildings in vegetation-dense environments may be partially occluded, making structural boundaries difficult to isolate from RGB imagery alone. Conventional augmentation strategies—such as rotation, flipping, or color jittering—modify existing images but do not address these limitations; they may even degrade critical structural cues by reducing edge sharpness or amplifying noise [25].

To explicitly address these limitations, we employ a fusion-based input enhancement strategy that enriches standard RGB inputs with auxiliary structural channels, increasing the information available to the model without requiring additional training samples [25]. Unlike instance-based augmentation, which operates on image-level transformations, Fusion Augmentation encodes domain-relevant structural information directly into the input representation, preserving the features most critical for distinguishing damage levels across diverse environmental conditions. To achieve this, we use three auxiliary representations: Canny edge detection to emphasize building boundaries and structural discontinuities; contrast enhancement to improve local visibility of subtle damage patterns; and unsharp masking to sharpen fine structural details such as cracks, roof degradation, and boundary irregularities.

https://doi.org/10.3390/rs18101529

3.3.1. Canny Edge Detection

Structural edges are critical for AI-based damage detection, as damage manifests through geometric discontinuities. Edges play a crucial role in characterizing building layouts, especially in post-disaster images, where environmental factors may obscure these features [39]. In this study, we employ the Canny detector [40] due to its robustness and proven effectiveness in extracting well-defined structural boundaries. The algorithm consists of the following stages:

1. Noise reduction: A Gaussian filter is applied to suppress high-frequency noise while preserving key structural edges intact. The Gaussian kernel is defined as

$$G(x,y) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{x^2+y^2}{2\sigma^2}\right) \tag{2}$$

where we use a 5×5 Gaussian kernel with $\sigma = 1.4$.

2. Gradient calculation: Image gradients G_x and G_y are computed using Sobel operators. The gradient magnitude and orientation are then derived as

$$G_{\text{edge}} = \sqrt{G_x^2 + G_y^2} \tag{3}$$

$$\theta = \tan^{-1}\left(\frac{G_y}{G_x}\right) \tag{4}$$

3. Non-maximum suppression: To produce thin, well-localized edges, non-maximum suppression retains only pixels whose gradient magnitude is a local maximum along the gradient direction.

4. Double thresholding: Edge pixels are classified using high (T_{high}) and low (T_{low}) thresholds. We set $T_{high} = 100$ and $T_{low} = 50$ based on validation experiments.

5. Edge tracking by hysteresis: Weak edges that are connected to the strong edges are retained, whereas isolated weak edges are suppressed. These steps remove noise while guaranteeing edge continuity.

The resulting edge map $E(x,y) \in \{0,1\}$ highlights building boundaries, structural discontinuities, and damage-related features. Integrating these edge maps into the network enables an early emphasis on geometric cues essential for identifying building damage.

3.3.2. Contrast Enhancement

Contrast enhancement is applied to improve the visibility of subtle damage features by redistributing pixel intensities across the full dynamic range. In this study, we apply Contrast Limited Adaptive Histogram Equalization (CLAHE) [41], which performs contrast enhancement locally with small image tiles of 8×8 pixels, while limiting the amplification of histogram bins through a clip limit of 2.0. This prevents the excessive noise enhancement that often occurs using global histogram equalization. CLAHE is applied independently to each RGB channel; the redistribution of histogram values is controlled by the following formulation:

$$H'(i) = \min\left(H(i), \frac{N_{\text{pixels}}}{N_{\text{bins}}} \cdot \text{clip\_limit}\right) \tag{5}$$

where $H(i)$ is the histogram count for bin (i). The resulting enhanced image $I_{enhanced}$ reveals finer texture variations, weathering patterns, and subtle structural differences that help distinguish visually altered yet structurally intact surfaces, such as soot or ash deposits, from genuine damage.

https://doi.org/10.3390/rs18101529

3.3.3. Unsharp Masking

Unsharp masking [42] is applied to emphasize fine structural details by sharpening the image through the combination of the original and a blurred version. The sharpened output is computed as

$$I_{sharp} = I + \alpha \cdot (I - G_\sigma * I) \tag{6}$$

where $G_\sigma * I$ denotes the Gaussian-blurred image using a 5×5 kernel with $\sigma = 1.0$, and the sharpening strength is set to $\alpha = 1.5$. This enhancement improves the visibility of subtle structural cues, such as small cracks and missing roof elements, while avoiding the artifacts that conventional sharpening filters may introduce.

3.3.4. Channel Fusion Strategy

After generating the auxiliary representations, including the edge map, contrast-enhanced image, and unsharp-masked image, each auxiliary output is converted into a single-channel grayscale representation and combined with the original RGB image. The fused input is constructed as

$$I_{fused} = [I_{RGB}, E, C, S] \tag{7}$$

where I_{RGB} denotes the three-channel RGB image, and E, C, and S denote the edge, contrast, and sharpness channels, respectively. This results in a six-channel input comprising three RGB channels and three auxiliary structural enhancement channels. The resulting representation preserves the visual content of the original image while providing additional information about boundaries, fine structural details, and local contrast that may not be clearly separable in RGB space alone. Using single-channel auxiliary features keeps the design compact while still capturing the structural cues relevant to damage assessment. Thus, the six-channel design balances richer input features with computational efficiency.

3.4. Domain Adaptation Techniques

Domain adaptation addresses the challenge of transferring knowledge learned from a source domain to a target domain with different underlying data distribution [43]. This capability is especially important in building damage assessment, where models trained on specific disasters and geographic regions must generalize to new and unseen scenarios. In this study, we investigate two complementary strategies: supervised fine-tuning, which depends on a small amount of labeled target data; and CORAL [44], an unsupervised feature alignment approach.

3.4.1. Supervised Fine-Tuning

Supervised fine-tuning adapts pretrained models to target domains using a limited number of labeled samples [45]. In this study, we use 30 labeled target-domain samples as a fixed, limited-label setting to examine whether a small amount of target supervision can improve transfer under geographic shift. In this approach, the early encode layers remain frozen to preserve low-level, transferable features such as edges and textures, while the deep encode layers and decoders are updated to learn domain-specific building characteristics and damage patterns. This selective updating allows the model to retain generalizable representations while effectively adapting to structural and environmental variations in the target domain. The fine-tuning process follows a standard training loop. Let M_θ be the pretrained model and $D_t = \{(x_i, y_i)\}_{i=1}^{N_t}$ the labeled target dataset. The early encoder layers are kept fixed, while the final encoder blocks and the decoder are unfrozen. For each

https://doi.org/10.3390/rs18101529

training batch (X_b, Y_b), predictions $\hat{Y}_b = M_\theta(X_b)$ are computed, and the parameters of the unfrozen layers are updated to minimize the composite loss.

$$L_{\text{total}} = \lambda_1 L_{\text{Dice}} + \lambda_2 L_{\text{Focal}} + \lambda_3 L_{\text{CE}},$$

with weights $\lambda_1 = 0.4$, $\lambda_2 = 0.4$, and $\lambda_3 = 0.2$, following the top-performing xView2 formulation [44]. In the present study, this composite loss formulation was adopted from the xView2 implementation, and the loss weights were kept fixed rather than re-optimized for the six-channel input setting. The model was optimized using AdamW for 10 epochs with a learning rate of 0.001. Through this procedure, the model gradually aligns its higher-level representations with the target distribution while preserving useful pretrained structure.

3.4.2. Unsupervised CORAL Domain Adaptation

Unsupervised domain adaptation is performed using CORrelation ALignment (CORAL) [44], which aligns second-order statistics between the source and target domains by matching their feature covariance matrices. CORAL minimizes the distributional discrepancy through the covariance-matching objective

$$L_{\text{CORAL}} = \frac{1}{4d^2} \|C_S - C_T\|_F^2, \tag{8}$$

where C_S and C_T denote the covariance matrices of the source and target features, respectively, and $\|\cdot\|_F$ is the Frobenius norm. The covariance matrices are computed as

$$C_{\text{S}} = \frac{1}{n_{\text{S}} - 1} \left(D_{\text{S}}^\top D_{\text{S}} - \frac{1}{n_{\text{S}}} (\mathbf{1}^\top D_{\text{S}})^\top (\mathbf{1}^\top D_{\text{S}}) \right), \tag{9}$$

$$C_{\text{T}} = \frac{1}{n_{\text{T}} - 1} \left(D_{\text{T}}^\top D_{\text{T}} - \frac{1}{n_{\text{T}}} (\mathbf{1}^\top D_{\text{T}})^\top (\mathbf{1}^\top D_{\text{T}}) \right). \tag{10}$$

where $D_S = \{f_i^s\}_{i=1}^{n_s}$ and $D_T = \{f_j^t\}_{j=1}^{n_t}$ are the source and target feature sets, respectively. To improve adaptation, CORAL loss is applied at multiple network layers, and the total multi-layer objective is

$$L_{\text{CORAL}}^{\text{total}} = \sum_{i=1}^{t} \lambda_i L_{\text{CORAL}}^{(i)}. \tag{11}$$

3.5. Benchmark Systems

In this paper, we used three xView2 competition winners as benchmark systems. The first-place xView2 solution [46] employs a combination of UNet-like architectures for building localization and Siamese neural networks for damage classification. Localization models are trained on pre-disaster images using several encoder backbones (e.g., ResNet34, SE-ResNeXt50, SENet154), which are then converted into Siamese networks for analyzing paired pre- and post-disaster inputs. The approach incorporates extensive data augmentation and a composite loss consisting of Dice, Focal, and cross-entropy terms.

The second-place solution [47] adopts UNet-like networks with pretrained DPN92 and DenseNet161 encoders for binary localization. Mixed-precision training (Apex) is used to improve computational efficiency, and localization and classification stages are optimized separately. Multiclass damage classification is performed using a Focal-with-Dice loss.

The third-place system [48] ensembles multiple semantic segmentation models trained with weighted cross-entropy to mitigate class imbalance. A shared encoder processes pre- and post-disaster imagery, whose extracted features are concatenated before entering the decoder. The ensemble spans several encoder families, including ResNets, DenseNets,

https://doi.org/10.3390/rs18101529

and EfficientNets paired with UNet or FPN decoders, and integrates pseudo-labeling and weighted model averaging to enhance robustness. All three benchmark systems report F1 scores above 0.85 on the official xView2 test set, reflecting the effectiveness of ensemble modeling and advanced training strategies. However, as demonstrated in our experiments, these approaches degrade substantially when evaluated on geographically out-of-distribution regions, motivating the development of our Fusion Augmentation and domain adaptation methods.

3.6. Evaluation Metrics

Model performance is assessed using metrics that capture both localization accuracy and damage classification quality, consistent with the official xView2 evaluation protocol.

F1 score. The primary metric is the F1 score [49], defined as the harmonic mean of precision and recall:

$$F1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{12}$$

where

$$\text{Precision} = \frac{TP}{TP + FP}, \qquad \text{Recall} = \frac{TP}{TP + FN}. \tag{13}$$

We compute F1 scores separately for localization and for each damage category (no damage, minor, major, and destroyed). Following the xView2 protocol, overall damage-classification performance is reported as the harmonic mean of the class-wise F1 scores:

$$\text{F1}_{\text{damage}} = \frac{4}{\sum_{c \in C} \frac{1}{\text{F1}_c}} \tag{14}$$

where $C = \{\text{no damage, minor, major, destroyed}\}$ ensuring that all categories contribute equally despite the dataset imbalance.

4. Results and Discussion

This section presents the experimental results and discusses the findings in the context of the existing literature.

4.1. Baseline Performance

The top three winning solutions from the xView2 challenge [46–48] represent the current state of the art in automated building damage assessment from satellite imagery. Table 3 presents the performance of these models on the official competition test set, demonstrating overall capabilities with localization F1 scores exceeding 0.84 and particularly high accuracy on the "no damage" class (F1 > 0.90). However, these models exhibit substantial performance degradation across two damage categories: minor and major. This pattern aligns with findings from the original xBD dataset paper [5].

Table 3. Performance of top three winning solutions on the xView2 Challenge test set adapted from the reported first-, second-, and third-place xView2 winning solutions [46–48].

Class	First Place	Second Place	Third Place
Localization	0.8621	0.8532	0.8465
No damage	0.9147	0.9018	0.9071
Minor damage	0.6385	0.6181	0.6173
Major damage	0.7819	0.7702	0.7651
Destroyed	0.8542	0.8486	0.8462

https://doi.org/10.3390/rs18101529

4.2. Geographic Bias

To investigate geographic biases, we constructed 17 location-specific test sets by withholding 10% of images from each disaster location in the xBD dataset. This approach allows a systematic evaluation of model generalization across diverse geographic contexts, disaster types, and environmental conditions. Tables 4–6 present comprehensive performance results for the top three winning solutions across all 17 xBD disaster locations and reveal substantial variation across regions and damage categories. A clear pattern emerges in several temperate U.S. regions, including Arkansas, Alabama, Florida, Missouri, Oklahoma, South Carolina, and Texas, where the models achieve localization F1 values above 0.87 and perform well across all damage categories. Several factors contributed to this achievement. Most of the training data in the xBD dataset comes from these locations, accounting for about 60% [5]. Second, they have relatively similar terrain, vegetation density, consistent building materials (mostly wood-frame construction with asphalt roofing), and architectural layouts, creating a homogeneous feature space that models can reliably learn. Third, the dominant disaster types in these regions, such as hurricanes, floods, and tornadoes, exhibit damage patterns well represented in the training corpus. This confirms Weber and Kané's findings [50] that the similarity of training data greatly impacts model performance in remote sensing applications.

Table 4. Location-wise performance of the top solution across 17 xBD disaster locations. Per-class F1 scores (localization, no damage, minor, major, destroyed) reveal significant geographic variability, forming the basis for subsequent geographic bias analysis [33,48] on each location from the xBD dataset.

Location	Disaster Type	Localization	No Damage	Minor	Major	Destroyed
Arkansas	Flood	0.8615	0.8962	0.7669	0.8882	0.0370
Alabama	Wind	0.8801	0.9448	0.7485	0.8796	0.9125
California	Fire	0.8689	0.8548	0.5572	0.2965	0.8934
Mexico	Earthquake	0.8769	0.9469	0.0000	0.127	0.0000
Florida	Wind	0.9112	0.9303	0.7444	0.8796	0.7668
Hawaii	Volcano	0.8790	0.8694	0.1202	0.0000	0.8161
Missouri	Wind	0.9405	0.9694	0.7170	0.8571	0.9358
Nepal	Flood	0.8375	0.8856	0.3036	0.8692	0.2060
Oklahoma	Wind	0.9416	0.8772	0.8054	0.8435	0.8983
Portugal	Fire	0.8391	0.8827	0.4691	0.3509	0.6328
South Australia	Fire	0.8357	0.8486	0.4257	0.0678	0.7188
South Carolina	Flood	0.8757	0.8667	0.6696	0.8270	0.1086
Indonesia	Tsunami	0.8741	0.9408	0.0000	0.6483	0.8649
Texas	Flood	0.9004	0.9246	0.7087	0.7518	0.297
Guatemala	Volcano	0.9089	0.8800	0.0000	0.0000	0.1264
Sunda	Tsunami	0.8653	0.8334	0.0000	0.0000	0.0000
Ayiti	Wind	0.8628	0.9380	0.5150	0.3694	0.5378

https://doi.org/10.3390/rs18101529

Table 5. Performance of second place solution [33,48] on each location from the xBD Dataset.

Location	Disaster Type	Localization	No Damage	Minor	Major	Destroyed
Arkansas	Flood	0.8285	0.8607	0.749	0.8665	0.0362
Alabama	Wind	0.8621	0.9221	0.7255	0.8581	0.8782
California	Fire	0.8453	0.8231	0.5454	0.2877	0.8651
Mexico	Earthquake	0.8465	0.9237	0.0000	0.1234	0.0000
Florida	Wind	0.8878	0.9073	0.7188	0.8567	0.7365
Hawaii	Volcano	0.8555	0.8426	0.1165	0.0000	0.7877
Missouri	Wind	0.9152	0.9446	0.6923	0.8342	0.9015
Nepal	Flood	0.8089	0.8564	0.2943	0.8446	0.1994
Oklahoma	Wind	0.9169	0.8521	0.7788	0.8212	0.8663
Portugal	Fire	0.8143	0.8547	0.4549	0.3406	0.6121
South Australia	Fire	0.8113	0.8225	0.4126	0.0658	0.6954
South Carolina	Flood	0.8501	0.8594	0.6487	0.8036	0.1052
Indonesia	Tsunami	0.8489	0.9163	0.0000	0.6292	0.8357
Texas	Flood	0.8756	0.8988	0.6866	0.7307	0.2877
Guatemala	Volcano	0.8833	0.8533	0.0000	0.0000	0.1223
Sunda	Tsunami	0.8401	0.8076	0.0000	0.0000	0.0000
Ayiti	Wind	0.8383	0.9122	0.4989	0.3585	0.5204

Table 6. Performance of third place solution [33,48] on each location from the xBD dataset.

Location	Disaster Type	Localization	No Damage	Minor	Major	Destroyed
Arkansas	Flood	0.8124	0.8412	0.7312	0.8451	0.0355
Alabama	Wind	0.8489	0.9087	0.7089	0.8396	0.8564
California	Fire	0.8312	0.8089	0.5334	0.2811	0.8445
Mexico	Earthquake	0.8321	0.9102	0.0000	0.1205	0.0000
Florida	Wind	0.8743	0.8934	0.7021	0.8378	0.7201
Hawaii	Volcano	0.8412	0.8289	0.1138	0.0000	0.7689
Missouri	Wind	0.9023	0.9312	0.6767	0.8156	0.8801
Nepal	Flood	0.7956	0.8421	0.2878	0.8267	0.1945
Oklahoma	Wind	0.9034	0.8389	0.7612	0.8045	0.8478
Portugal	Fire	0.8012	0.8401	0.4445	0.3329	0.5978
South Australia	Fire	0.7978	0.8089	0.4034	0.0643	0.6789
South Carolina	Flood	0.8367	0.8445	0.6334	0.7856	0.1028
Indonesia	Tsunami	0.8334	0.9012	0.0000	0.6156	0.8167
Texas	Flood	0.8623	0.8834	0.6712	0.7145	0.2811
Guatemala	Volcano	0.8644	0.8540	0.0000	0.0000	0.1222
Sunda	Tsunami	0.8213	0.7978	0.0000	0.0000	0.0000
Ayiti	Wind	0.8161	0.8989	0.4876	0.3543	0.5108

In contrast to temperate regions, Mediterranean and tropical regions (Hawaii, Portugal, South Australia, Indonesia, Ayiti) show modest performance degradation, with F1 scores typically 10–15% lower. The training distribution does not account for environmental variability in these places. For instance, Hawaii's volcanic landscapes feature dense tropical vegetation and unique geological formations that obscure the boundaries of buildings. The Mediterranean climate of Portugal affects sunlight, vegetation (eucalyptus forests), and construction materials (stone and concrete) differently than North American training data. High humidity, dense vegetation, and informal settlement patterns, combined with non-Western architectural styles, characterize Indonesia's tropical climate. Environmental variability leads to domain shift [51], where visual characteristics from temperate training data may not transfer well to climatically diverse test conditions. Vegetation and other non-building objects can introduce false alarms and reduce the accuracy of building detection in high-resolution satellite imagery [52].

https://doi.org/10.3390/rs18101529

To better understand whether regional performance differences are primarily explained by training sample quantity, Figure 4 compares the number of training images per location with the corresponding macro-F1 score. No clear relationship emerges: Portugal and South Australia, despite having among the largest training sets in the benchmark, achieve macro-F1 scores substantially below locations such as Missouri and Oklahoma, which are trained on comparatively fewer samples. Conversely, several low-sample locations outperform high-sample counterparts on the same metric.

The Nepal case illustrates this more clearly. Nepal does not have the largest training sample in the dataset. However, it still contains a substantial number of training images and exhibits weak performance for certain damage categories, particularly minor damage (F1 = 0.304) and Destroyed (F1 = 0.206), while major damage remains comparatively strong (F1 = 0.869). This class-dependent behavior suggests that transfer performance cannot be attributed solely to limited sample availability. Instead, it likely reflects a combination of geographic and structural mismatch, including terrain complexity, shadow effects, dense settlement patterns, and local architectural characteristics that differ from the dominant patterns in the training data.

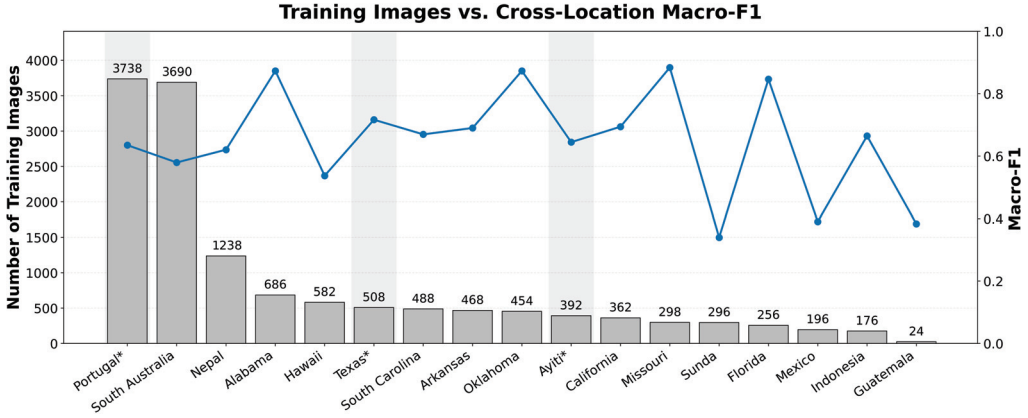


Figure 4. Relationship between training sample count and macro-F1 across xBD locations. Bars show the number of training images, and the line shows the corresponding macro-F1 score. Locations marked with * denote out-of-domain test locations. The results show that having more training images does not necessarily correspond to better cross-location generalization.

Similar behavior can be observed in other low-performing regions, as shown in Tables 4–6. For example, Mexico and Guatemala show near-zero or zero F1 scores for some damage categories, while Sunda also performs poorly across multiple classes. To provide qualitative insight into this behavior, Figure 5 shows a representative example from Guatemala comparing localization results before and after Fusion Augmentation. These cases likely reflect a combination of environmental variation, hazard-specific damage appearance, class imbalance, and reduced class separability due to domain shift, rather than a simple shortage of training examples. Overall, the results indicate that cross-location degradation in building damage assessment is shaped not only by how much data is available, but also by how well the source data captures the structural and environmental characteristics of the target region.

https://doi.org/10.3390/rs18101529



(**a**) (**b**) (**c**)

Figure 5. Representative qualitative comparison from Guatemala. Panel (**a**) shows a post-disaster satellite image obtained from the publicly available xBD dataset [5], used under the Creative Commons license stated by the dataset source. Panels (**b**) and (**c**) show localization results before and after Fusion Augmentation, respectively, generated by the authors using the proposed method. Green regions in subfigures (**b**) and (**c**) indicate segmented building areas. Red circles were added by the authors to highlight representative regions where additional building structures are recovered after augmentation.

Geographic bias has major effects on practical disaster response. Despite strong aggregate performance metrics, state-of-the-art models may perform poorly in settings with environmental or structural differences from the training data. A model trained on North American disasters may have 90% accuracy in Florida but only 30% accuracy in Nepal, making it unreliable for worldwide humanitarian applications. The variety in dependability defies the universality assumption implicit in many catastrophe informatics systems and highlights the necessity for domain adaptation strategies that explicitly address geographic bias.

4.3. Domain Adaptation to Unseen Geographic Regions

To rigorously quantify geographic bias and evaluate mitigation strategies, we designed controlled out-of-domain experiments using three representative locations—Texas (temperate), Ayiti (tropical), and Portugal (Mediterranean)—that were completely excluded from model training. This experimental design simulates realistic deployment scenarios, where emergency responders must assess damage in regions for which pretrained models have no direct experience. In this setting, the domain adaptation analysis is used to examine how representative adaptation strategies affect cross-location performance under geographic shifts in the xBD benchmark. For each target location, we trained the first-place winning solution [46] from scratch on the remaining 16 xBD locations, creating truly out-of-domain test conditions. We then evaluated six configurations: (i) pretrained baseline model, (ii) pretrained model with Fusion Augmentation applied during training, (iii) pretrained model with supervised fine-tuning on 30 labeled target samples, (iv) pretrained model with both fine-tuning and Fusion Augmentation, (v) pretrained model with unsupervised CORAL domain adaptation [44], and (vi) pretrained model with CORAL and Fusion Augmentation. This comprehensive evaluation enables isolation of individual technique contributions and assessment of synergistic effects. The results, presented in Tables 7–9, reveal nuanced patterns of adaptation effectiveness across geographic contexts.

https://doi.org/10.3390/rs18101529

Table 7. Domain adaptation results for Texas (temperate region). SDA: supervised domain adaptation, FA: Fusion Augmentation, and UDA: unsupervised domain adaptation. Bold values indicate the best performance for each metric.

Method	Localization	No Damage	Minor	Major	Destroyed
Pretrained model [46]	0.8236	0.6673	0.4105	0.5528	0.2141
Pretrained model+FA	0.8285	0.6631	0.4344	0.5721	0.2154
Pretrained model+SDA fine-tuning	0.8429	0.6742	0.4723	0.6037	0.2931
Pretrained model+SDA+FA	**0.8485**	**0.6961**	**0.5047**	**0.6320**	**0.3117**
Pretrained model+UDA-CORAL	0.8246	0.6581	0.4553	0.4446	0.2259
Pretrained model+UDA-CORAL+FA	0.8315	0.6641	0.4616	0.4494	0.2631

Table 7 reveals moderate domain shift in climatically similar but geographically diverse locations. Under these conditions of moderate domain-shift due to geographic specificity, the pretrained baseline localization performance (F1 = 0.824) is acceptable, but intermediate damage categories (minor: 0.411, major: 0.553) show significant degradation (33% and 19%, respectively) compared to in-domain performance on similar climatic disasters. This degradation, despite Texas being a temperate flood scenario similar to training data from Arkansas and South Carolina, reveals that geographic specificity—urban development patterns, building codes, construction practices—introduces sufficient feature distribution shift to impair model performance. Fusion Augmentation alone shows slight improvements (minor: +5.8%, major: +3.5%), indicating that explicit structural feature enhancement can withstand domain transfer without target-specific adaptation.

Table 8. Domain adaptation results for Ayiti (tropical region). SDA: supervised domain adaptation, FA: Fusion Augmentation, and UDA: unsupervised domain adaptation. Bold values indicate the best performance for each metric.

Method	Localization	No Damage	Minor	Major	Destroyed
Pretrained model [46]	0.6852	0.6214	0.3108	0.1892	0.2046
Pretrained model+FA	0.7043	0.6638	0.3916	0.2631	0.2354
Pretrained model+SDA fine-tuning	0.8124	0.6946	0.4315	0.2981	0.2852
Pretrained model+SDA+Augmentation	**0.8413**	**0.6961**	**0.4958**	**0.3419**	**0.3140**
Pretrained model+UDA-CORAL	0.7246	0.6417	0.4054	0.2531	0.2419
Pretrained model+UDA-CORAL+FA	0.7515	0.6511	0.4212	0.3058	0.2913

Supervised fine-tuning has a significantly greater impact (minor: +15.0%, major: +9.2%), highlighting the significance of target-domain supervision. Combining fine-tuning with Fusion Augmentation yields the highest results (minor: 0.505, major: 0.632), with 23% and 14% improvements above baseline. For Texas, unsupervised CORAL domain adaptation [44] yields mixed results, improving minor damage (+10.9%) but reducing major damage performance (−19.6%). A possible explanation is negative transfer: although CORAL reduces the global feature distribution mismatch between the source and target domains, second-order alignment may also weaken class-discriminative boundaries when target-domain damage patterns are not well aligned with those of the source domain. In the Texas case, this may have reduced the separability of the major damage class, leading to degraded performance despite moderate gains in other categories. This suggests that supervised adaptation with limited labeled target data is more effective than unsupervised feature alignment under modest domain shift.

Table 8 shows more significant domain shift effects in tropical conditions. The pretrained baseline severely fails under domain shift in tropical conditions characterized by dense vegetation and environmental occlusion (localization: 0.685, minor: 0.311, major: 0.189), indicating that temperate disaster models are unsuitable for tropical environments. Tropical deployments demonstrate a significant geographic bias, with a 60% decrease in minor damage performance and a 76% decrease in major damage performance compared

https://doi.org/10.3390/rs18101529

to temperate regions. Ji et al. [53] found that tropical vegetation presents unique challenges for building detection, including dense canopy cover, confusing shadows, and moisture-dependent reflectance changes, hindering damage assessment. Fusion Augmentation significantly improves performance (minor: +26.0%, major: +39.0%), indicating that edge detection and contrast enhancement partially compensate for environmental occlusion by highlighting structural boundaries despite vegetation covering RGB features. Combined fine-tuning with Fusion Augmentation yields the best tropical performance (localization: 0.841, minor: 0.496, major: 0.342), with 59% and 81% gains over the baseline. In significant-domain-shift scenarios, unsupervised CORAL adaptation shows noteworthy outcomes (minor: +30.4%, major: +33.8%), indicating that statistical feature distribution alignment can be beneficial even without target labels. CORAL + Fusion Augmentation enhances unsupervised adaptation (minor: +35.5%, major: +61.7%), comparable to supervised fine-tuning without labeled target data. This has crucial implications for rapid natural-disaster response, where labeled data may not be available during the initial evaluation.

Table 9. Domain adaptation results for Portugal (Mediterranean region). SDA: supervised domain adaptation, FA: Fusion Augmentation, and UDA: unsupervised domain adaptation. Bold values indicate the best performance for each metric.

Method	Localization	No Damage	Minor	Major	Destroyed
Pretrained model [46]	0.7049	0.6415	0.3049	0.1992	0.3985
Pretrained model+FA	0.7251	0.6832	0.3257	0.2118	0.4157
Pretrained model+SDA fine-tuning	0.8028	0.7212	0.4082	0.3046	0.4453
Pretrained model+SDA+FA	**0.8285**	**0.7516**	**0.4267**	**0.3685**	**0.4875**
Pretrained model+UDA-CORAL	0.7564	0.6762	0.3481	0.2478	0.4295
Pretrained model+UDA-CORAL+FA	0.7917	0.6839	0.3795	0.3495	0.4495

Table 9 shows that Portugal presents a distinct domain shift scenario combining Mediterranean climate characteristics with wildfire-specific damage patterns. Even though Portugal's Mediterranean environment is closer to tropical training regions, intermediate damage detection dropped to tropical Ayiti levels (localization: 0.705, minor: 0.305, major: 0.199). A number of factors cause this poor generalization. Portugal's eucalyptus forests have dense, uniform vegetation canopies with spectral signatures distinct from those of temperate deciduous forests and tropical rainforests in the training data. Ash covering from the 2017 Portuguese wildfires can mask building boundaries and create false damage signatures, as ash-covered roofs may resemble collapsed structures [54]. Additionally, Portuguese construction predominantly features stone and concrete, with terracotta roofing, and architectural characteristics that diverge from wood-frame North American structures, creating fundamental feature mismatches. Fusion Augmentation yields slightly smaller increases (minor: +6.8%, major: +6.3%) than Ayiti, possibly due to ash coverage affecting edge-recognition techniques. Supervised fine-tuning yields significant improvements (minor: +33.9%, major: +52.9%), whereas the combined technique achieves the best performance (minor: 0.427, major: 0.369), representing 40% and 85% over the baseline, respectively. CORAL + Fusion Augmentation enhances unsupervised performance (minor: +24.5%, major: +75.4%), with CORAL adaptation showing significant benefits (minor: +14.2%, major: +24.4%).

A number of patterns show up across all three out-of-domain locations, indicating geographic bias and adaptation techniques. As shown in Table 10, the proposed method achieves consistent improvements across all three locations. First, baseline performance degradation is proportional to environmental dissimilarity rather than geographic distance. Ayiti and Portugal experience severe degradation despite being geographically and climatically distinct, whereas Texas experiences moderate degradation despite being far from training regions. This shows that domain shift in disaster damage assessment is

https://doi.org/10.3390/rs18101529

environmental and structural rather than geographical. Fusion Augmentation has constant gains (5–40% depending on shift severity) across every location, demonstrating the fundamental robustness of specific structural feature encoding to environmental variations. Next, fine-tuning and Fusion Augmentation work together to provide complementary advantages, resulting in 23–85% enhancements based on damage category and location. Fusion Augmentation yields superior structural representations that fine-tuning can more effectively adjust to target-domain characteristics, possibly due to complementary mechanisms. Finally, unsupervised CORAL adaptation shows variable effectiveness, with limited benefit for moderate domain shifts (Texas) but substantial gains for extreme shifts (Ayiti, Portugal), suggesting that statistical feature alignment becomes more valuable as shift severity increases and label distributions diverge.

Table 10. Performance improvement for each damage class across out-of-domain locations using the top solution. We report the pretrained baseline, the proposed method (pretrained model + supervised domain adaptation + fusion augmentation), absolute improvement (Δ), and relative improvement (Δ%).

Class	Texas				Ayiti				Portugal			
	Baseline	Proposed	Δ	Δ%	Baseline	Proposed	Δ	Δ%	Baseline	Proposed	Δ	Δ%
No damage	0.6673	0.6961	0.0288	4.3	0.6214	0.6961	0.0747	12.0	0.6415	0.7516	0.1101	17.2
Minor	0.4105	0.5047	0.0942	22.9	0.3108	0.4958	0.1850	59.5	0.3049	0.4267	0.1218	39.9
Major	0.5528	0.6320	0.0792	14.3	0.1892	0.3419	0.1527	80.7	0.1992	0.3685	0.1693	85.0
Destroyed	0.2141	0.3117	0.0976	45.6	0.2046	0.3140	0.1094	53.5	0.3985	0.4875	0.0890	22.3

This study has several limitations that should be considered when interpreting the findings. First, and most significantly, all domain adaptation experiments are conducted within the xBD benchmark; the absence of external validation on an independent post-disaster dataset limits the generalizability of the reported improvements beyond this benchmark. Second, the unsupervised adaptation setting evaluates CORAL with a fixed low-resource budget of 30 labeled samples, without systematic comparison against more recent domain adaptation methods, such as DANN [51], or adversarial feature alignment approaches. Third, loss-weight configurations (λ_1, λ_2, λ_3) were adopted from the baseline xView2 formulation without re-optimization for the six-channel input setting; a dedicated sensitivity analysis of alternative weighting schemes remains an important direction for future work.

5. Conclusions

In this study, we investigate generalization challenges and geographic biases in AI-based building damage assessment from satellite imagery. Through extensive evaluation of the top-performing xView2 solutions across 17 disaster locations, we found that geographic and structural characteristics, as well as insufficient training data, degrade model performance. Our results demonstrate that structural feature enhancement through auxiliary channels is not merely beneficial but essential for robust AI-based damage detection across geographically diverse disaster scenarios. The Nepal case, together with the broader location-wise analysis, shows that environmental context, architectural diversity, and terrain complexity impact generalization behavior, and that training sample quantity alone does not fully explain regional performance differences. The experimental results show a 7.1% increase in overall F1 score, with significant gains in intermediate damage categories like minor and major damage. Domain adaptation experiments on three unseen locations demonstrate that Fusion Augmentation with supervised fine-tuning improves the minor and major classes by 40.8% and 60.0%, respectively, while unsupervised CORAL improves the minor and major damage classes by 24.2% and 39.5% over the benchmarks.

https://doi.org/10.3390/rs18101529

Future research could explore several directions to further improve cross-domain generalization in building damage assessment. First, incorporating foundation models, such as large-scale vision transformers pretrained on diverse remote sensing datasets, could help learn more transferable features and reduce the need for domain-specific fine-tuning. In addition, generative approaches like GAN-based domain adaptation (e.g., CycleGAN) could be used to synthesize realistic target-domain images from source data, improving performance under strong domain shifts. Finally, exploring more diverse data augmentation strategies and domain adaptation techniques on larger and more varied datasets could further enhance model robustness and generalization.

Author Contributions: Conceptualization, S.K., R.B. and A.D.; data curation, B.C. and U.T.; formal analysis, B.C. and U.T.; funding acquisition, S.K., R.B. and A.D.; investigation, B.C.; methodology, S.K., R.B. and A.D.; project administration, S.K., R.B. and A.D.; resources, S.K., R.B. and A.D.; software, B.C. and U.T.; supervision, S.K., R.B. and A.D.; validation, B.C. and U.T.; visualization, B.C. and U.T.; writing—original draft, B.C. and U.T.; writing—review and editing, S.K., U.T. and R.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Kansas NSF EPSCoR grant (R55156).

Data Availability Statement: The xBD dataset used in this study is publicly available for download through the xView2 Challenge website: https://xview2.org/ (accessed on 1 January 2023). Further dataset details are available via the DOI record: https://doi.org/10.1184/R1/8135576.v1.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. NASA. What Are the Effects of Climate Change on Extreme Weather? 2024. Available online: https://science.nasa.gov/climate-change/extreme-weather/ (accessed on 26 November 2024).
2. United Nations Office for Disaster Risk Reduction (UNDRR). Hazard Definition and Classification Review: Technical Report. 2024. Available online: https://www.undrr.org/publication/hazard-definition-and-classification-review-technical-report (accessed on 26 November 2024).
3. Defense Innovation Unit (DIU). Assessing Building Damage from Satellite Imagery. 2024. Available online: https://www.diu.mil/ai-xview-challenge (accessed on 26 November 2024).
4. xView2 Challenge Team. Computer Vision for Building Damage Assessment. 2024. Available online: https://xview2.org/ (accessed on 26 November 2024).
5. Gupta, R.; Goodman, B.; Patel, N.; Hosfelt, R.; Sajeev, S.; Heim, E.; Doshi, J.; Lucas, K.; Choset, H.; Gaston, M. Creating xBD: A dataset for assessing building damage from satellite imagery. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–17 June 2019; pp. 10–17.
6. Melamed, Z.; Shahar, Y.; Waisglass, N.; Kliger, M.; Averbuch-Elor, H. Bias in Building Damage Detection from Satellite Imagery. *arXiv* **2021**, arXiv:2104.06533.
7. Wiguna, S.; Adriano, B.; Mas, E.; Koshimura, S. Evaluation of deep learning models for building damage mapping in emergency response settings. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 5651–5667. [CrossRef]
8. Ryan-Mosley, T. How AI Can Actually Be Helpful in Disaster Response. 2023. Available online: https://www.technologyreview.com/2023/02/20/1068824/ai-actually-helpful-disaster-response-turkey-syria-earthquake/ (accessed on 26 November 2024).
9. Duarte, D.; Nex, F.; Kerle, N.; Vosselman, G. Satellite image classification of building damages using airborne and satellite image samples in a deep learning approach. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2018**, *4*, 89–96. [CrossRef]
10. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
11. Wu, C.; Zhang, F.; Xia, J.; Xu, Y.; Li, G.; Xie, J.; Du, Z.; Liu, R. Building damage detection using U-Net with attention mechanism from pre- and post-disaster remote sensing datasets. *Remote Sens.* **2021**, *13*, 905. [CrossRef]
12. Dong, L.; Shan, J. A comprehensive review of earthquake-induced building damage detection with remote sensing techniques. *ISPRS J. Photogramm. Remote Sens.* **2013**, *84*, 85–99. [CrossRef]
13. Kim, D.; Won, J.; Lee, E.; Park, K.R.; Kim, J.; Park, S.; Yang, H.; Cha, M. Disaster assessment using computer vision and satellite imagery: Applications in detecting water-related building damages. *Front. Environ. Sci.* **2022**, *10*, 969758. [CrossRef]

14. Zheng, Z.; Zhong, Y.; Wang, J.; Ma, A.; Zhang, L. Building damage assessment for rapid disaster response with a deep object-based semantic change detection framework: From natural disasters to man-made disasters. *Remote Sens. Environ.* **2021**, *265*, 112636. [CrossRef]
15. Kaur, N.; Lee, C.C.; Mostafavi, A.; Mahdavi-Amiri, A. Large-scale building damage assessment using a novel hierarchical transformer architecture on satellite images. *Comput.-Aided Civ. Infrastruct. Eng.* **2023**, *38*, 2072–2091. [CrossRef]
16. Tsai, F.J.; Lin, S.Y. A class distance penalty deep learning method for post-disaster building damage assessment. *KSCE J. Civ. Eng.* **2024**, *28*, 2005–2019. [CrossRef]
17. Jakubik, J.; Muszynski, M.; Vössing, M.; Kühl, N.; Brunschwiler, T. Toward foundation models for earth monitoring: Generalizable deep learning models for natural hazard segmentation. In Proceedings of the IGARSS 2023—2023 IEEE International Geoscience and Remote Sensing Symposium, Pasadena, CA, USA, 16–21 July 2023; pp. 5638–5641.
18. Ahn, K.; Han, S.; Park, S.; Kim, J.; Park, S.; Cha, M. Generalizable disaster damage assessment via change detection with vision foundation model. In Proceedings of the AAAI Conference on Artificial Intelligence, Philadelphia, PA, USA, 25 February–4 March 2025; Volume 39, pp. 27784–27792.
19. Zhao, F.; Zhang, C.; Zhang, R.; Wang, T. Visual prompt learning of foundation models for post-disaster damage evaluation. *Remote Sens.* **2025**, *17*, 1664. [CrossRef]
20. Yu, B.; Sun, Y.; Hu, J.; Chen, F.; Wang, L. Post-disaster building damage assessment based on gated adaptive multi-scale spatial-frequency fusion network. *Int. J. Appl. Earth Obs. Geoinf.* **2025**, *141*, 104629. [CrossRef]
21. Melamed, D.; Johnson, C.; Gerg, I.D.; Zhao, C.; Blue, R.; Hoogs, A.; Clipp, B.; Morrone, P. Uncovering bias in building damage assessment from satellite imagery. In Proceedings of the IGARSS 2024—2024 IEEE International Geoscience and Remote Sensing Symposium, Athens, Greece, 7–12 July 2024; pp. 8095–8099.
22. Gevaert, C.M.; Buunk, T.; van den Homberg, M.J.C. Auditing geospatial datasets for biases: Using global building datasets for disaster risk management. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 12579–12590. [CrossRef]
23. Adedeji, O.; Owoade, P.; Ajayi, O. Image augmentation for satellite images. *arXiv* **2022**, arXiv:2207.14580. [CrossRef]
24. Hao, X.; Liu, L.; Yang, R. A review of data augmentation methods of remote sensing image target recognition. *Remote Sens.* **2023**, *15*, 827–867. [CrossRef]
25. Ghaffar, M.A.A.; McKinstry, A.; Maul, T.; Vu, T.T. Data augmentation approaches for satellite image super-resolution. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2019**, *4*, 47–54. [CrossRef]
26. Takahashi, R.; Matsubara, T.; Uehara, K. Data augmentation using random image cropping and patching for deep CNNs. *IEEE Trans. Circuits Syst. Video Technol.* **2019**, *30*, 2917–2931. [CrossRef]
27. Buslaev, A.; Iglovikov, V.I.; Khvedchenya, E.; Parinov, A.; Druzhinin, M.; Kalinin, A.A. Albumentations: Fast and flexible image augmentations. *Information* **2020**, *11*, 125. [CrossRef]
28. DeVries, T. Improved regularization of convolutional neural networks with Cutout. *arXiv* **2017**, arXiv:1708.04552. [CrossRef]
29. Kshirsagar, S.R.; Falk, T.H. Quality-aware bag of modulation spectrum features for robust speech emotion recognition. *IEEE Trans. Affect. Comput.* **2022**, *4*, 1892–1905. [CrossRef]
30. Kshirsagar, S.; Falk, T.H. Cross-language speech emotion recognition using bag-of-word representations, domain adaptation, and data augmentation. *Sensors* **2022**, *22*, 6445. [CrossRef] [PubMed]
31. Kshirsagar, S.; Pendyala, A.; Falk, T.H. Task-specific speech enhancement and data augmentation for improved multimodal emotion recognition under noisy conditions. *Front. Comput. Sci.* **2023**, *5*, 1039261. [CrossRef]
32. Parupati, B.C.R.; Kshirsagar, S.; Bagai, R.; Dutta, A. Towards robust building damage detection: Leveraging augmentation and domain adaptation. In Proceedings of the 2025 IEEE Green Technologies Conference (GreenTech), Wichita, Kansas, 26–28 March 2025; pp. 163–167.
33. Parupati, B.C.R. Enhancing Generalizability in Building Damage Detection: Domain Adaptation and Augmentation Approaches for Post-Disaster Assessment. Ph.D. Thesis, Wichita State University, Wichita, KS, USA, 2024.
34. Xu, Q.; Shi, Y.; Yuan, X.; Zhu, X.X. Universal domain adaptation for remote sensing image scene classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 4700515. [CrossRef]
35. Hu, Y.; Tang, H. On the generalization ability of a global model for rapid building mapping from heterogeneous satellite images of multiple natural disaster scenarios. *Remote Sens.* **2021**, *13*, 984–1008. [CrossRef]
36. Kshirsagar, S.R. *Affective Human-Machine Interfaces: Towards Multi-lingual, Environment-Robust Emotion Detection from Speech*; Institut National de la Recherche Scientifique: Quebec, QC, Canada, 2022.
37. Mouradi, A.; Kshirsagar, S. Robust Building Damage Detection in Cross-Disaster Settings Using Domain Adaptation. *arXiv* **2026**, arXiv:2603.14694. [CrossRef]
38. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: New York ,NY, USA, 2009; pp. 248–255.
39. Yang, S.; Peng, B.; Sui, T.; Wu, M.; Huang, Q. Advancing Self-Supervised Learning for Building Change Detection and Damage Assessment: Unified Denoising Autoencoder and Contrastive Learning Framework. *Remote Sens.* **2025**, *17*, 2717. [CrossRef]

https://doi.org/10.3390/rs18101529

40. Canny, J. A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **1986**, *PAMI-8*, 679–698. [CrossRef]
41. Zuiderveld, K. Contrast limited adaptive histogram equalization. In *Graphics Gems*; Academic Press Professional, Inc.: San Diego, CA, USA, 1994; pp. 474–485.
42. Polesel, A.; Ramponi, G.; Mathews, V.J. Image enhancement via adaptive unsharp masking. *IEEE Trans. Image Process.* **2000**, *9*, 505–510. [CrossRef] [PubMed]
43. Wilson, G.; Cook, D.J. A survey of unsupervised deep domain adaptation. *ACM Trans. Intell. Syst. Technol. (TIST)* **2020**, *11*, 1–46. [CrossRef]
44. Sun, B.; Saenko, K. Deep CORAL: Correlation alignment for deep domain adaptation. In Proceedings of the Computer Vision—ECCV 2016 Workshops, Amsterdam, The Netherlands, 8–10. 15–16. October 2016; pp. 443–450.
45. Yosinski, J.; Clune, J.; Bengio, Y.; Lipson, H. How transferable are features in deep neural networks? In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 3320–3328.
46. Durnov, V. xView2 First Place Solution. 2019. Available online: https://github.com/vdurnov/xview2_1st_place_solution (accessed on 10 September 2024).
47. Seferbekov, S. xView2 Second Place Solution. 2019. Available online: https://github.com/selimsef/xview2_solution (accessed on 10 September 2024).
48. Khvedchenya, E. xView2 Third Place. 2020. Available online: https://github.com/DIUx-xView/xView2_third_place (accessed on 10 September 2024).
49. Powers, D.M. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv* **2020**, arXiv:2010.16061. [CrossRef]
50. Weber, E.; Kané, H. Building Disaster Damage Assessment in Satellite Imagery with Multi-Temporal Fusion. *arXiv* **2020**, arXiv:2004.05525. [CrossRef]
51. Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; Lempitsky, V. Domain-Adversarial Training of Neural Networks. *J. Mach. Learn. Res.* **2016**, *17*, 1–35.
52. You, Y.; Wang, S.; Ma, Y.; Chen, G.; Wang, B.; Shen, M.; Liu, W. Building Detection from VHR Remote Sensing Imagery Based on the Morphological Building Index. *Remote Sens.* **2018**, *10*, 1287. [CrossRef]
53. Ji, S.; Wei, S.; Lu, M. Building Extraction in Tropical Regions from High-Resolution Satellite Imagery Using Multispectral and Shadow Information. *Remote Sens.* **2019**, *11*, 2005. [CrossRef]
54. Silva, P.R.D.; Pereira, J.C.; Moreira, F.; Oliveira, T.M. Wildfire Damage Assessment in Portugal Using High-Resolution Satellite Imagery. *Remote Sens.* **2018**, *10*, 1792. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

https://doi.org/10.3390/rs18101529

fire

MDPI

Article

Nationwide Daily Wildfire Occurrence Prediction Using Time Proxy Variables and the Canadian Fire Weather Index (FWI)

Boksoo Choi and Gye-Young Kim *

Department of AI·SW Convergence, Soongsil University, Seoul 06978, Republic of Korea; cbs1988@soongsil.ac.kr
* Correspondence: gykim11@ssu.ac.kr

Abstract

Climate change has intensified global wildfire risks, yet national-scale prediction remains challenging due to the difficulty of consistently monitoring fuel conditions and human ignition factors. This study introduces calendar-based time proxy variables as structural surrogates for these unobservable drivers and integrates them with the Canadian Fire Weather Index (FWI) within a parsimonious framework for seasonally fire-prone regions such as South Korea. Using 15 years of nationwide wildfire records and daily observations from 100 ASOS stations (2011–2025), predictive performance was evaluated across eight models and five feature sets (Time-only, Weather-only, Weather + Time, FWI-only, and FWI + Time). Based on test-set mean AUC, the Time-only feature set reached 0.7374, clearly exceeding the random-classifier baseline (AUC = 0.5) and indicating the independent predictive value of time proxy variables. Furthermore, integrating time proxies with FWI improved performance, with the best model (CatBoost) achieving test AUC = 0.8394 and Recall = 0.6019. Multi-model SHAP analysis revealed complementary contributions of FWI components (53.7% ± 4.7%) and time proxy variables (46.3% ± 4.7%). Overall, the results demonstrate that a simple yet structured input design based on time proxy variables provides meaningful predictive performance for nationwide wildfire early warning systems.

Keywords: wildfire occurrence prediction; Fire Weather Index (FWI); time proxy variables; parsimonious prediction model; early warning system; SHAP

1. Introduction

Rising temperatures, altered precipitation patterns, and prolonged drying associated with climate change have increased wildfire risk worldwide, making wildfires an increasingly important hazard across diverse climatic zones [1–3]. Accordingly, the need for predictive models that enable early detection and proactive management of wildfire risk has grown substantially. Although wildfire occurrence is a complex phenomenon shaped by the interaction of multiple environmental and human-related factors [4–6], a major challenge in predictive modeling lies in the difficulty of consistently observing and quantifying fuel conditions and human activity factors at the national scale [7–9]. To address this limitation, this study examines whether calendar-based temporal variables can serve as structural proxies that indirectly capture fuel conditions and human behavioral characteristics that are difficult to observe directly. Based on this idea, we propose a parsimonious wildfire prediction framework that integrates time proxy variables with the Canadian Fire Weather Index (FWI) and evaluate its predictive performance using nationwide wildfire occurrence data.

1.1. Wildfire Occurrence Mechanisms

Wildfire occurrence is commonly explained as the interaction of three essential elements—weather, fuel, and ignition source—often conceptualized as a fire-regime triangle [10,11]. These elements jointly determine whether a wildfire event can occur and how it spreads once ignited. Among these elements, weather conditions play a dominant role in determining wildfire risk. Temperature, relative humidity, wind speed, and precipitation influence fuel moisture conditions and directly affect both ignition probability and fire spread dynamics [11]. Figure 1 illustrates this fire-regime triangle, in which wildfire occurrence is understood as the overlapping outcome of weather conditions, fuel availability, and ignition sources.

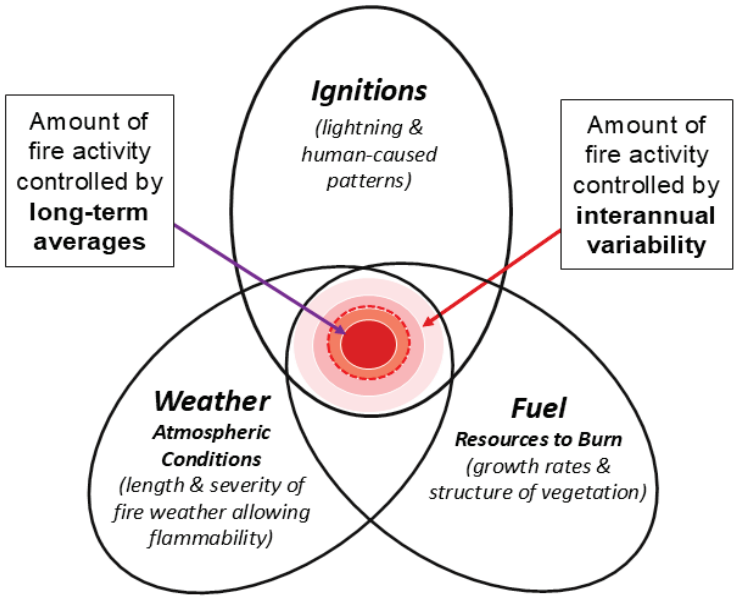


Figure 1. Fire regime triangle illustrating the interaction of weather, fuel, and ignition sources in wildfire occurrence. Wildfire events occur when all three components are simultaneously present. The colored arrows and shaded circles conceptually indicate the relative influence of long-term averages and interannual variability on wildfire activity. Source: Adapted from Moritz et al. [12].

To integrate these meteorological factors, the Canadian Fire Weather Index (FWI) system has been widely used as a comprehensive indicator of wildfire danger [4,13]. The FWI system sequentially calculates fuel moisture and fire behavior indicators based on meteorological observations and consists of several sub-indices, including the Fine Fuel Moisture Code (FFMC), Duff Moisture Code (DMC), Drought Code (DC), Initial Spread Index (ISI), and Buildup Index (BUI) [4,13,14]. Due to its ability to represent meteorological influences on wildfire behavior, the FWI has been widely adopted in operational wildfire monitoring and forecasting systems [14,15].

Fuel conditions represent another critical component of wildfire occurrence. Fuel refers to the amount, spatial continuity, and moisture content of combustible materials such as leaf litter, dead wood, understory vegetation, and forest canopy fuels [6,7]. Fuel accumulation and drying processes evolve gradually through seasonal climatic conditions, and fuel moisture serves as an important threshold for wildfire ignition. However, fuel characteristics vary across vegetation types and landscapes, making it difficult to consistently observe and quantify them as nationwide datasets [7].

Ignition sources constitute the third component of wildfire occurrence mechanisms. Many wildfire events are caused by human activities rather than natural processes such as lightning. Previous studies show that human activities account for the majority of

https://doi.org/10.3390/fire9060217

wildfire ignitions in fire-prone regions, with values reaching approximately 84% in the conterminous United States [16] and similarly high proportions in parts of Europe [5,17]. Human-caused ignition is closely linked to socio-economic activities, including outdoor recreation, agricultural burning, and daily human activities near forested areas. However, these behavioral factors are rarely available as structured datasets at national scales.

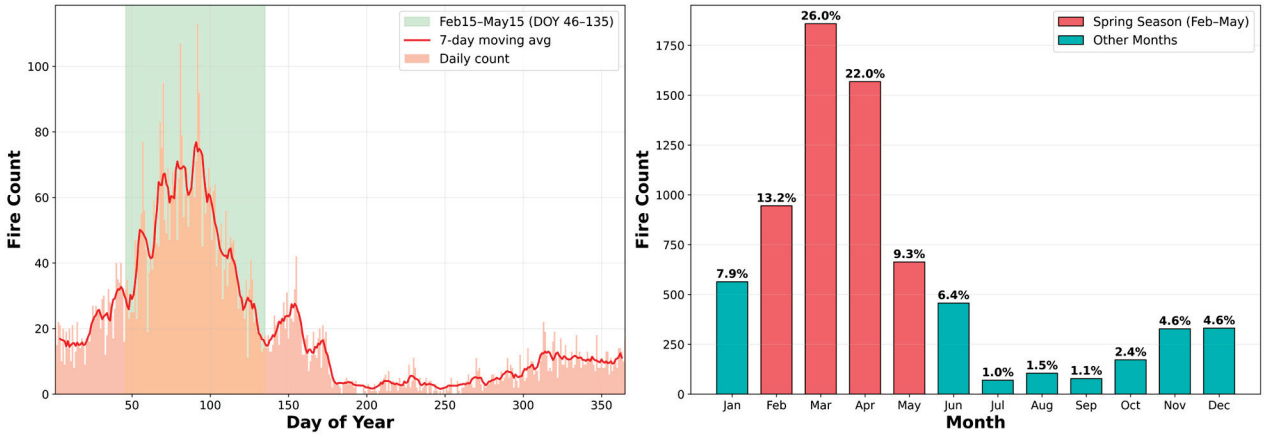
Although satellite-derived vegetation indices, vegetation status, live fuel moisture, land-cover data, and other remote-sensing products can provide valuable information on fuel conditions, their operational use often requires additional preprocessing, spatial–temporal harmonization, and continuous national-scale availability. Human behavioral factors are also rarely available as structured datasets at national scales. These constraints motivate the search for parsimonious approaches that can indirectly represent latent fuel and ignition-related patterns in predictive models.

1.2. Temporal Regularity of Wildfire Occurrence

Wildfire occurrence is not randomly distributed but exhibits distinct temporal regularities, concentrating within specific timeframes [16–18]. In regions with strong seasonality, such as Korea, wildfire occurrence is typically concentrated in spring. Within a day, wildfire occurrence is more frequent during daytime hours (approximately 11:00–17:00), and human-caused wildfires tend to increase during weekends, public holidays, and seasonal outdoor activity periods such as grave-visiting or leisure seasons [16,18].

Figure 2 shows that these temporal regularities are also evident in the Korean wildfire records used in this study. Wildfire occurrence varies systematically with the day of year, month, day of week, and hour of day. These patterns arise from the combined effects of seasonal meteorological variation, fuel accumulation and drying processes, and cyclical human activities [16–18]. Accordingly, wildfire occurrence can be interpreted as the outcome of a time-structured socio-ecological system [17].

In regions with strong seasonality, such temporal patterns tend to recur consistently over time. This suggests that temporal information itself contains meaningful signals related to wildfire occurrence probability [16,17]. Therefore, wildfire occurrence should not be explained solely by instantaneous meteorological conditions but rather understood within a broader temporal context [17].



(**a**) (**b**)

Figure 2. *Cont.*

https://doi.org/10.3390/fire9060217

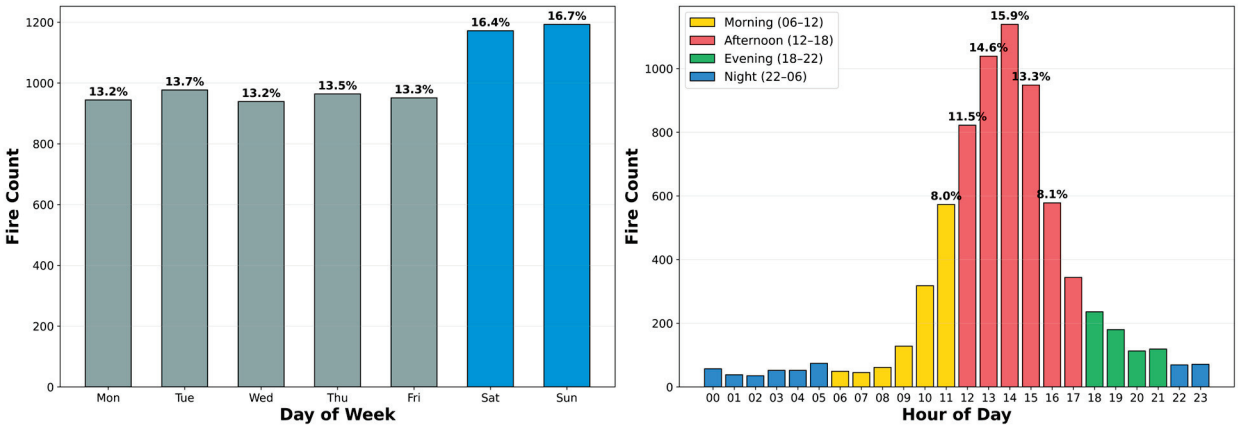


Figure 2. Relationship between time and wildfire occurrence (2011–2025). (**a**) Wildfire occurrence by day of year (**b**) Wildfire occurrence by month. (**c**) Wildfire occurrence day of week (**d**) Wildfire occurrence by hour of day.

1.3. Time Proxy Approach

Wildfire occurrence arises from the interaction of meteorological conditions, fuel accumulation and drying processes, and human-caused ignition. However, unlike meteorological variables, fuel conditions and human behavior are difficult to observe and quantify consistently at the national scale and are unevenly distributed across space and time [6]. To address this limitation, this study adopts time proxies as a structural alternative that indirectly represents the temporal variability of these unobserved factors. This approximation does not imply that fuel conditions or human behavior are determined solely by calendar time; rather, time proxies serve as structural surrogates for recurrent seasonal and social rhythms, while their ability to capture non-stationary changes in vegetation condition, live fuel moisture, or ignition environments under climate change remains limited. The use of such variables has been discussed as a way to represent temporal structures indirectly in time series modeling [19].

A time proxy refers to the use of calendar-based temporal variables to capture variation in factors that are difficult to measure directly. Such variables—including season, month, day of year, day of week, and time of day—have been widely used as proxies for latent states or behaviors in various domains, including economics, transportation, epidemiology, and energy demand forecasting [19]. In wildfire prediction, time proxies can serve as surrogate variables that reflect the temporal dynamics of fuel conditions and human activities [20]. Fuel accumulation and drying evolve gradually with seasonal climate patterns, while human activities follow socially structured temporal cycles such as weekdays, weekends, holidays, and agricultural periods [6,18].

Accordingly, time proxies can be interpreted not merely as auxiliary variables but as structural covariates that encode the shared temporal structure of fuel conditions and human behavior. This perspective is particularly relevant in regions such as Korea, where climatic seasonality is pronounced and socio-economic activities are strongly organized around calendar cycles. From a practical standpoint, time proxies offer key advantages: they can be consistently defined across all locations nationwide, require minimal data construction, and can be readily implemented in real-time operational systems. These properties make them especially suitable for developing parsimonious and scalable wildfire prediction models at the national scale.

While recent machine learning–based wildfire studies have incorporated temporal variables to account for seasonality [8,20], they have mostly treated them as control terms

https://doi.org/10.3390/fire9060217

rather than as structural predictors representing latent drivers of wildfire occurrence. In contrast, this study conceptualizes time proxies as structural surrogate variables and empirically evaluates their effectiveness in representing fuel and human behavioral processes.

1.4. Research Objectives and Hypotheses

The objective of this study is to examine whether time proxy variables can serve as structural predictors of wildfire occurrence and to develop a parsimonious prediction framework that integrates temporal and meteorological information at the national scale. Wildfires are inherently rare and highly uncertain events. In practical wildfire management, providing reliable early warning signals that support resource allocation and preparedness may be more valuable than achieving extremely high spatial precision [6]. Under this perspective, this study explores whether a prediction framework based on widely available meteorological and temporal variables can provide robust predictive performance for nationwide wildfire monitoring.

Specifically, this study evaluates whether time proxy variables can serve as effective predictors of wildfire occurrence both independently and in combination with meteorological variables. Based on this objective, the following research hypotheses (RHs) are examined:

RH1. *A model composed solely of time proxy variables (Time-only) provides significantly higher predictive performance than random prediction.*

RH2. *Integrating time proxy variables with meteorological variables or the Fire Weather Index (FWI) significantly improves predictive performance compared with models that rely solely on meteorological variables or FWI.*

The main contributions of this study lie in reframing temporal variables as structural predictors of wildfire occurrence and demonstrating their effectiveness within a parsimonious prediction framework. They are summarized as follows:

1. Introducing a time-proxy-based wildfire occurrence prediction framework in which calendar-based temporal variables indirectly represent latent fuel conditions and human activity patterns.
2. Evaluating the independent and complementary predictive roles of time proxy variables when combined with meteorological and FWI variables.
3. Demonstrating that a parsimonious model consisting of 11 variables can achieve robust predictive performance for nationwide wildfire occurrence prediction.
4. Providing interpretable insights into the complementary roles of FWI components and time proxy variables through multi-model SHAP analysis.

Unlike many region-specific wildfire prediction models that require high-dimensional spatial or socio-economic datasets, the proposed framework provides a simple, scalable, and operationally feasible approach for nationwide wildfire prediction and early warning systems.

The remainder of this paper is organized as follows. Section 2 describes the datasets and prediction methods used in this paper. Section 3 presents the experimental results. Section 4 discusses the implications and limitations of the proposed approach. Finally, Section 5 concludes the paper and outlines directions for future research.

https://doi.org/10.3390/fire9060217

2. Materials and Methods

2.1. Data Sources

This study utilized two primary datasets: nationwide wildfire occurrence records and meteorological observations.

Wildfire occurrence data were obtained from the national wildfire statistics database provided by the Korea Forest Service. As shown in Table 1, the dataset covers the period from 2011 to 2025 and includes detailed information on wildfire events such as occurrence date, location, and administrative region. A total of 7140 wildfire events were initially recorded during the study period. The collected variables include the wildfire occurrence time (start_time), suppression completion time (end_time), occurrence location (province and municipality), burned area (damagearea, ha), cause of ignition (firecause), and suppression duration (duration_min), among others.

Table 1. Wildfire data.

Item	Description
Data source	Korea Forest Service wildfire statistics
Data period	2011–2025 (15 years)
Total number of occurrences	7140
Unit of analysis	517,448 (station-days)

The meteorological data were collected from the daily observation records of the Automated Synoptic Observing System (ASOS) operated by the Korea Meteorological Administration. ASOS is an automated system that provides meteorological observations of uniform quality across the country and offers consistent long-term time-series data.

As shown in Table 2, this study used daily meteorological observations from 100 ASOS weather stations. During the observation period, a total of 549,231 daily meteorological records (station × day) were collected. The collected variables include 36 items, covering both basic information, such as station and date, and observed meteorological variables, such as mean temperature, maximum temperature, minimum temperature, mean relative humidity, minimum relative humidity, daily precipitation, maximum hourly precipitation, mean wind speed, maximum wind speed, and total solar radiation.

Table 2. Meteorological data.

Item	Description
Data source	Korea Meteorological Administration Open MET Data Portal, ASOS
Collection period	2010–2025
Study area	South Korea, covered by 100 ASOS stations
Weather stations/observation unit	100 ASOS stations/station-day records
Number of records (station-days)	Total 549,231; 2011–2025 (517,448) (2010: 31,783, used only as a warm-up period for FWI calculation)
Meteorological information	36 variables

Note: Meteorological observations were retained as ASOS station-day records and were not spatially averaged or interpolated to produce a nationwide gridded map.

The meteorological data were used both as input variables and for the calculation of the FWI. Because FWI is recursively calculated by accumulating the previous day's

https://doi.org/10.3390/fire9060217

index values, time is required during the initial spin-up (warm-up) period for the values to stabilize under the initial conditions [4]. Therefore, in this study, the 31,783 records from 2010 (station-days) were used only as warm-up data for FWI calculation and initial fuel moisture stabilization, whereas the actual training and evaluation of the prediction models were conducted using 517,448 records (station-days) from 2011 to 2025.

2.2. Input Variable Construction

2.2.1. Weather Variables

In this study, among the 36 meteorological variables observed by ASOS, four variables used in FWI calculation—daily mean temperature (Temperature), daily mean relative humidity (Relative Humidity), daily mean wind speed (Wind_speed), and daily precipitation (Rainfall)—were selected as the basic input variables, as shown in Table 3. This selection was made to ensure consistency with the FWI-based feature sets, enable a fair comparison between raw weather variables and derived FWI components, and maintain a parsimonious and operationally feasible model structure.

Table 3. Weather-only input variable configuration.

Variable Name	Description
Temperature	Daily mean temperature (°C)
Relative Humidity	Daily mean relative humidity (%)
Wind speed	Daily mean wind speed (m/s)
Precipitation	Daily precipitation (mm)

Figure 3 shows that these four meteorological variables exhibit clear monthly patterns. These variables are key meteorological factors that directly affect wildfire occurrence and were also used as the reference feature set for performance comparison with the FWI-based approach.

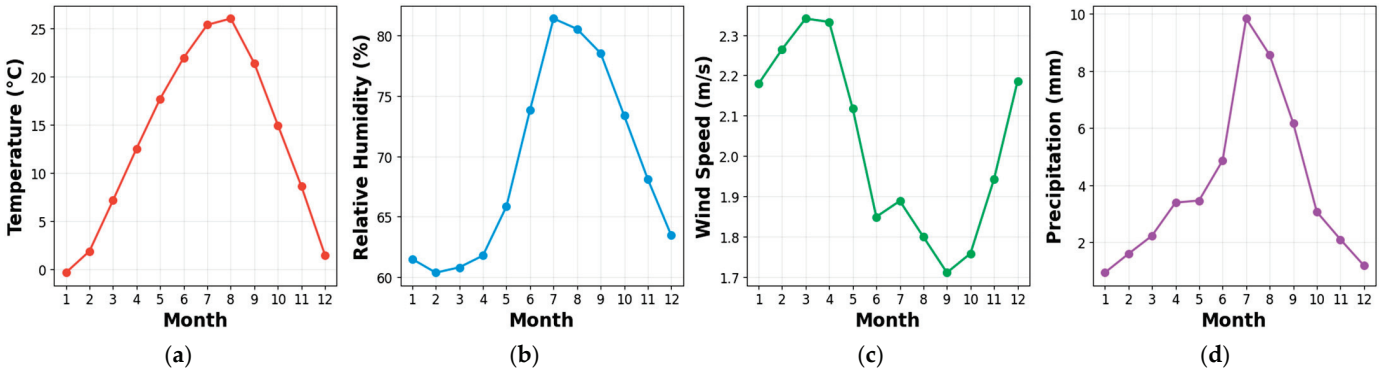


Figure 3. Monthly average weather patterns (2010–2025). (**a**) Temperature, (**b**) Relative humidity, (**c**) Wind speed, (**d**) Precipitation.

2.2.2. Fire Weather Index (FWI)

The Canadian Fire Weather Index (FWI) system was used as a core input variable in this study. It is a standard wildfire danger rating system that transforms daily meteorological variables into six sub-indices representing fuel moisture conditions and fire potential, and integrates them into a final fire danger index [4,13]. As summarized in Table 4, the system consists of FFMC, DMC, DC, ISI, BUI, and FWI, and its recursive calculation structure reflects cumulative drying and wetting processes over time [4].

https://doi.org/10.3390/fire9060217

Table 4. Fire Weather Index (FWI) components.

Index	Full English Name	Main Elements	Physical Meaning
FFMC	Fine Fuel Moisture Code	Temperature, relative humidity, wind speed, precipitation	Short-term moisture status of fine fuels
DMC	Duff Moisture Code	Temperature, relative humidity, precipitation	Dryness status of the organic layer
DC	Drought Code	Temperature, precipitation	Drought and cumulative dryness status of deep fuels
ISI	Initial Spread Index	FFMC, wind speed	Initial fire spread potential
BUI	Build Up Index	DMC, DC	Fuel accumulation and available combustible mass
FWI	Fire Weather Index	ISI, BUI	Overall wildfire danger level

In this study, FWI components were calculated according to the standard Canadian Fire Weather Index algorithms described by Van Wagner and Pickett [21] and Van Wagner [4]. The calculation was implemented in Python 3.12.13 within the Google Colab environment using the open-source code provided by Quilcaillé et al. [22], which follows the original Van Wagner formulation. The calculations were performed independently for each of the 100 ASOS stations in a station-wise sequential manner. Daily temperature (°C), relative humidity (%), wind speed (m/s, converted to km/h), and precipitation (mm) were used as inputs. Standard initial values were applied ($FFMC_0$ = 85.0, DMC_0 = 6.0, DC_0 = 15.0), and meteorological data from 2010 were used exclusively as a warm-up period for initial fuel moisture stabilization. The resulting FWI components for 2011–2025 were then used as predictor variables in the wildfire occurrence models. Implementation validity was checked against the reference test dataset provided in the cffdrs R package [23].

Table 5 summarizes the distribution of the six FWI components across 548,961 station-days, after excluding 270 records (0.05%) with missing meteorological inputs from the original 549,231 station-days. Overall, the indices showed substantial variability across stations and dates, reflecting heterogeneous fire-weather conditions. The percentile values further indicate that some components had skewed distributions. These excluded records contained missing values across all six components because of missing meteorological inputs and were omitted from subsequent modeling.

Table 5. FWI summary statistics.

Statistic	FFMC	DMC	DC	ISI	BUI	FWI
Count	548,961	548,961	548,961	548,961	548,961	548,961
Mean	67.58	9.31	65.03	1.68	12.35	2.14
Std	21.64	10.81	70.09	1.67	13.75	3.1
25%	55.5	2.07	10.44	0.42	2.68	0.16
50%	78.06	5.88	41.23	1.34	7.94	0.66
75%	83.66	12.54	96.89	2.52	17.17	3.15
Max	95.02	156.06	557.57	94.64	160.96	57.49

2.2.3. Time Proxy Variables

As shown in Figure 2, wildfire occurrence in Korea exhibits distinct seasonal and cyclical patterns according to date, month, and day of the week. Based on this pattern, and in line with the study objective of nationwide daily wildfire prediction and the principle

https://doi.org/10.3390/fire9060217

of parsimonious model design, this study selected a total of five time-proxy variables, as presented in Table 6.

Table 6. Time proxy variable configuration.

Category	Variable Name	Description
Cyclical variable	dayofyear_sin	Day of year (sine transformed)
Cyclical variable	dayofyear_cos	Day of year (cosine transformed)
Binary variable	is_spring	Spring season indicator (February–May)
Binary variable	is_autumn	Autumn season indicator (October–December)
Binary variable	is_weekend	Weekend indicator

First, the day of year was selected to reflect the fact that fuel conditions and human activity patterns vary over time. However, if cyclical and periodic variables such as day of year are used directly, an artificial discontinuity arises at the boundary between the end and the beginning of the year. One transformation method used to preserve the cyclical continuity of such temporal variables within the model is sine–cosine-based cyclical encoding. This method is a classical approach based on Fourier series and circular statistics theory, and it has been widely used in signal processing, statistical analysis, and machine learning to continuously represent periodic variations such as seasonality and weekly cycles [24,25].

Figure 4 visually illustrates the process of sine and cosine encoding. If only sine encoding is used, different points in time may have the same value (e.g., 30° and 150°). However, when cosine encoding is applied together, each point in time is mapped to a unique coordinate pair (x, y) in a circular space, thereby creating a continuous cyclical structure in which the beginning and end of time are not disconnected. Figure 4a shows the sine–cosine encoding results of the day of year, and Figure 4b presents its projection into circular space.

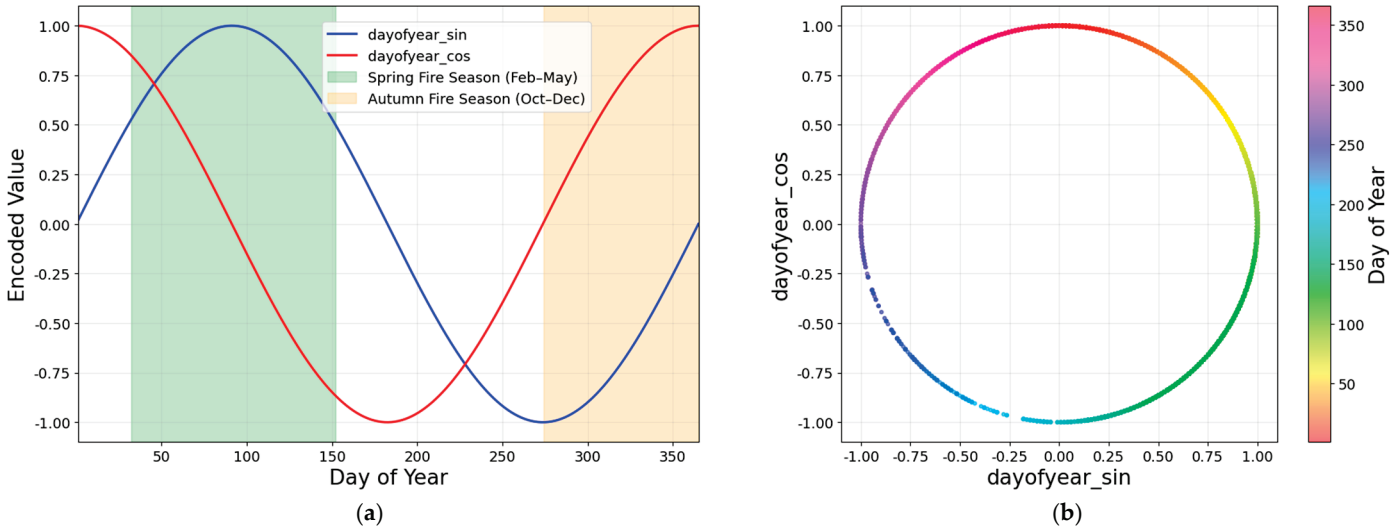


Figure 4. Transforming time proxy variables for wildfire prediction. (**a**) Cyclical encoding of day of year (**b**) Wildfire events in cyclical space.

https://doi.org/10.3390/fire9060217

When this sine–cosine-based cyclical encoding is applied, the day of year t is expressed for cycle length T as follows:

$$f_s(t) = \sin\left(\frac{2\pi t}{T}\right),\ f_c(t) = \cos\left(\frac{2\pi t}{T}\right)$$

Here, t ∈{0, . . ., T − 1}, and the cycle length T was set to 365 in common years and 366 in leap years. This transformation was intended to eliminate the artificial discontinuity occurring at the boundary between the end and the beginning of the year and to preserve the cyclical continuity of the temporal variable within the model. In addition, although the cyclical variable month of year, which also reflects annual seasonality, was considered as a time proxy variable, it was excluded because of its high informational overlap with day of year.

Next, in order to reflect seasonal asymmetry, dummy variables for spring (is_spring, February–May) and autumn (is_autumn, October–December) were added. This reflects the empirical characteristic that wildfire occurrence in Korea is concentrated in specific seasons [18]. By contrast, summer and winter were not included as separate seasonal dummy variables because their wildfire occurrence frequency is relatively low and their independent explanatory power was considered limited.

In addition, the weekend indicator (is_weekend) was included to reflect the structural difference in the intensity of human activity between weekdays and weekends [17]. Although the hour variable can explain intraday variation in wildfire occurrence, it was excluded in this study because the prediction was performed at the daily scale and daily meteorological data were used, and thus model consistency was considered. This configuration of time proxy variables was determined by simultaneously considering predictive performance, interpretability, and applicability at the national scale.

2.3. Data Processing

2.3.1. Wildfire–Meteorological Data Integration

In this study, spatial matching and temporal matching were sequentially performed to integrate wildfire occurrence data with meteorological observation data. Spatial matching was conducted by identifying the nearest meteorological station based on the latitude and longitude coordinates of each wildfire occurrence point. The KDTree algorithm, a binary tree structure for efficient nearest-neighbor search in multidimensional space, was applied to systematically identify candidate stations for each fire event [26]. To minimize data loss due to missing meteorological observations at the nearest station, a fallback matching algorithm that searches up to the 10th nearest station when data are unavailable at closer stations was implemented.

As shown in Table 7, this approach successfully matched all 7140 wildfire events to meteorological stations: 6702 events (93.9%) were matched to the nearest station, while 438 events (6.1%) were matched to fallback stations (2nd to 5th nearest). The mean distance between wildfire locations and matched stations was 15.82 km, with a maximum distance of 73.41 km. For modeling purposes, the dependent variable was defined at the station-day level as a binary indicator of wildfire occurrence (occurrence = 1, non-occurrence = 0). During the matching process, 501 wildfire events were found to overlap on the same station-day. Because the prediction unit of this study is the station-day, these overlapping events were consolidated into a single positive case. Consequently, the final target variable consisted of 6639 unique wildfire-occurrence station-days. This binary variable (fire_occurred) indicates whether at least one wildfire occurred within the jurisdiction of the corresponding observation station on a given date, and it was used as the target variable in the classification models.

https://doi.org/10.3390/fire9060217

Table 7. Wildfire–meteorological data matching results (2011–2025).

Item	Value
Total number of observations (station-days)	517,448
Number of wildfire events	7140 (100%)
Matched to nearest station	6702 (93.9%)
Matched to fallback stations (2nd–5th)	438 (6.1%)
Duplicate wildfire events on the same station-day	501
Wildfire occurrence (station-days)	6639
Wildfire Occurrence Rate (Wildfire-occurrence station-days/total station-days)	1.28%

Among the total 517,448 observations (station-days) from 2011 to 2025, wildfire occurrence (fire_occurred = 1) was observed in 6639 cases, accounting for 1.28% of the total. This shows that wildfire occurrence was a relatively rare event in the dataset and that appropriate handling of class imbalance was necessary in model construction [27].

2.3.2. Data Splitting and Class Imbalance Handling

To properly evaluate model performance, this study adopted a temporal split strategy. As shown in Table 8, the training data consisted of 8 years of data from 2011 to 2018, the validation data consisted of 3 years from 2019 to 2021, and the test data consisted of 4 years from 2022 to 2025. The validation data were used for hyperparameter tuning and threshold optimization, whereas the test data were used only for the final performance evaluation of the models in order to prevent overfitting.

Table 8. Data split and class imbalance handling.

Category	Period	Number of Observations	Wildfire Occurrence (A)	Wildfire Non-Occurrence (B)	Class Ratio (A:B)	Wildfire Occurrence Rate (%)
Train	2011–2018	273,405	3197	270,208	1:85	1.17
Train (After Sampling)	2011–2018	9591	3197	6394	1:2.0	33.33
Validation	2019–2021	103,902	1508	102,394	1:68	1.45
Testing	2022–2025	140,141	1934	138,207	1:71	1.38

Next, to address the severe class imbalance in which wildfire occurrences accounted for only 1.28% of all observations, the following strategy was applied. First, as shown in Table 8, year-wise balanced undersampling was performed on the training data (2011–2018), whereby non-occurrence cases equal to twice the number of wildfire occurrence cases in each year were randomly sampled to construct a 1:2 class ratio of wildfire occurrence to non-occurrence cases. Undersampling is a resampling technique that reduces the number of majority class instances to balance the class distribution; random undersampling, as employed here, randomly removes samples from the majority class while preserving all minority class instances [28]. Second, undersampling was not applied to the validation and test data; instead, the original class distribution (approximately 1.45% and 1.38%, respectively) was preserved so that prediction performance could be evaluated under actual operational conditions. Finally, to compensate for the potential information loss

https://doi.org/10.3390/fire9060217

caused by undersampling, additional class weights were applied in logistic regression to prevent underestimation of the minority class.

2.4. Research Design and Definition of Feature Sets

In order to systematically analyze performance differences according to input variable composition, this study defined five feature sets, as shown in Table 9.

Table 9. Composition and roles of feature sets.

Feature Set	Number of Variables	Composition
Time-only	5	dayofyear_sin/cos, is_spring, is_autumn, is_weekend
Weather-only	4	Temperature, Relative Humidity, Wind Speed, Precipitation
Weather + Time	9	Weather (4) + Time (5)
FWI-only	6	FFMC, DMC, DC, ISI, BUI, FWI
FWI + Time	11	FWI (6) + Time (5)

The Time-only feature set was used to test RH1 by examining whether time proxy variables alone show meaningful differences from the baseline. The Weather-only and FWI-only feature sets were used to test how much explanatory power is provided by weather variables and FWI variables alone, respectively [4]. The Weather + Time and FWI + Time feature sets were used to test RH2 by examining whether performance improves when time proxy variables are added to the Weather-only and FWI-only feature sets [18].

2.5. Prediction Models

We compared eight prediction models, spanning a statistical model, tree-based ensemble methods, and deep learning models, under identical data splits and feature sets. Logistic Regression served as the baseline because of its interpretability. To capture nonlinear relationships and feature interactions, we included Random Forest, Extra Trees, and gradient boosting models (GBM, XGBoost, LightGBM, and CatBoost).

Class imbalance was handled using class weights or equivalent model-specific parameters. Deep learning models (MLP and LSTM) were additionally included as benchmarks: MLP learns nonlinear patterns from tabular features, whereas LSTM is intended to capture temporal dependencies in sequential inputs. Through this comparison of machine learning and deep learning approaches, we examine whether parsimonious models can still achieve competitive predictive performance under limited input conditions. Specifically, for deep learning models, we evaluated different architectural complexities (e.g., Simple and Complex variants for LSTM) to ensure a robust comparison with machine learning models.

2.6. Performance Evaluation Metrics and Statistical Validation

As shown in Table 10, model performance was primarily evaluated using the area under the receiver operating characteristic curve (AUC), which measures the model's ability to discriminate between wildfire and non-wildfire events across all classification thresholds. Because wildfire occurrence is a rare event with a low occurrence rate relative to the total number of observations, Recall, Precision, and F1-score were additionally calculated as auxiliary metrics to better reflect detection performance for the positive class.

https://doi.org/10.3390/fire9060217

Table 10. Summary of performance evaluation metrics.

Indicators	Formula	Meaning
AUC	Area under the ROC curve	Overall classification performance
Recall	TP/(TP + FN)	Wildfire detection rate
Precision	TP/(TP + FP)	Alarm accuracy
F1-Score	$'2 \times (P \times R)/(P + R)$	Balance between precision and recall

Specifically, because failing to detect a wildfire (False Negative) entails significantly more severe consequences than a false alarm (False Positive), Recall was regarded as the most important performance metric. Through this, the study aimed to comprehensively evaluate not only overall discriminative ability but also how effectively actual wildfire occurrences could be detected without being missed.

To quantify the uncertainty of each performance metric, 95% confidence intervals were estimated using a non-parametric bootstrap method [29]. This approach reflects sample variability without relying on distributional assumptions and improves the stability of performance estimation in rare-event prediction problems. In addition, to determine whether performance differences between models were statistically significant, the DeLong test was applied to compare the AUCs of correlated ROC curves derived from the same test dataset [30]. The DeLong test is a non-parametric method for assessing the statistical significance of differences between AUC values while accounting for the fact that the predictions of the two models are based on the same observations.

2.7. Importance Analysis of Time Proxy Variables: SHAP Analysis

SHAP (SHapley Additive exPlanations) analysis was conducted to interpret the relative contributions of the FWI components and time proxy variables to prediction [31]. SHAP is a game-theoretic model interpretation method widely used to quantify variable importance in complex machine learning models.

In this study, SHAP values were used to assess the importance of individual variables and the group-level contribution of the FWI and time proxy variables. The analysis was applied to six tree-based ensemble models—Random Forest, Gradient Boosting, XGBoost, LightGBM, CatBoost, and Extra Trees—and the results were summarized across models to reduce dependence on any single algorithm. This step was intended to clarify whether time proxy variables provided substantial predictive information beyond the FWI-based inputs.

Figure 5 presents the overall conceptual framework of this study. It illustrates how wildfire occurrence was modeled using meteorological information, FWI-based fuel representation, and calendar-based time proxy variables reflecting recurrent temporal patterns in fuel conditions and human activity. It also summarizes the two main hypotheses tested in this study: the independent predictive value of time proxies (RH1) and their complementary effect when combined with FWI (RH2). All analyses were conducted in Google Colab using Python 3.12.13 and machine learning libraries including scikit-learn, XGBoost, LightGBM, CatBoost, and SHAP.

https://doi.org/10.3390/fire9060217

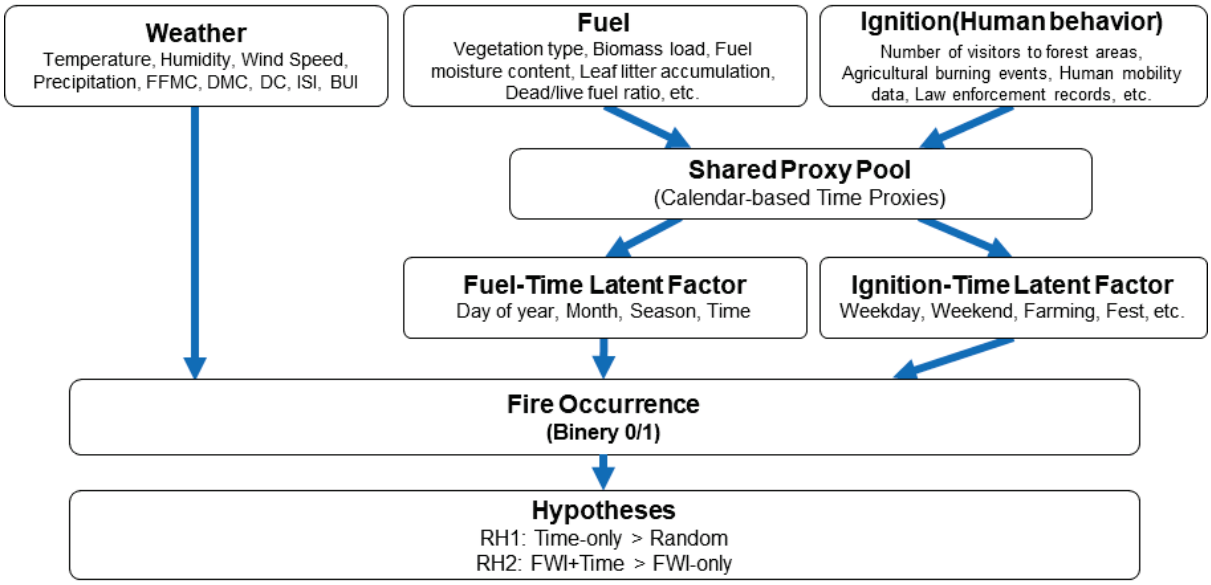


Figure 5. Conceptual framework of wildfire occurrence prediction using FWI and time proxy variables.

3. Results and Analysis

3.1. Model Comparison on the Validation Set

In this section, predictive performance for wildfire occurrence was compared across the five feature sets defined in Table 9 using the validation set, with the aim of selecting a reference model for subsequent analysis. This step was intended not to determine final generalization performance, but to assess the relative effects of feature composition and model choice. The feature-set definitions are provided in Table 9, whereas Table 11 summarizes the AUC results for each feature set and model on the validation set.

Table 11. Comparison of AUC across prediction models and feature sets (Validation Set).

Feature Set	LR	RF	GBM	XGB	LGBM	CatBoost	Extra Trees	MLP	Mean
Time-Only	0.7495	0.7120	0.7238	0.7242	0.7245	0.7392	0.7223	0.7459	0.7302
Weather-Only	0.7881	0.7966	0.7978	0.7995	0.7980	0.8033	0.8014	0.8014	0.7983
Weather + Time	0.8130	0.8068	0.8067	0.8091	0.8076	0.8158	0.8089	0.8170	0.8106
FWI-Only(A)	0.7958	0.8117	0.8091	0.8128	0.8119	0.8129	0.8027	0.8081	0.8081
FWI + Time(B)	0.8323	0.8380	0.8374	0.8406	0.8416	0.8411	0.8379	0.8410	0.8387
(B)-(A)	0.0365	0.0263	0.0283	0.0278	0.0297	0.0282	0.0352	0.0329	0.0306

Overall, performance differences among models within the same feature set were modest. This pattern was most evident for the FWI + Time feature set, where the AUC ranged from 0.8323 to 0.8416, yielding a narrow maximum-minimum gap of 0.0093. By contrast, differences across feature sets were more substantial, suggesting that predictive performance was influenced more strongly by feature composition than by the specific algorithm used.

The Time-only feature set, based solely on calendar-derived variables, produced AUC values between 0.7120 and 0.7495 (mean 0.7302), indicating meaningful but limited predictive power. When only meteorological information was used, the FWI-Only achieved a higher mean AUC (0.8081) than Weather-Only (0.7983), implying that FWI provides a more

https://doi.org/10.3390/fire9060217

informative representation of wildfire-related weather conditions than raw meteorological variables alone.

Adding time proxy variables improved performance in both meteorological settings. The mean AUC increased from 0.7983 to 0.8106 for Weather-Only → Weather + Time, and from 0.8081 to 0.8387 for FWI-Only → FWI + Time. The latter corresponds to a mean gain of 0.0306, indicating that time proxy variables contributed more strongly when combined with FWI than with raw weather variables. Under the FWI + Time feature set, LightGBM achieved the highest validation AUC (0.8416), followed by CatBoost (0.8411), MLP (0.8410), and XGBoost (0.8406). Although the AUC values across all eight models ranged from 0.8323 to 0.8416, the gap among the top four models was only 0.0010. These results suggest that several leading models yielded highly similar performance once the most informative feature set was applied.

Taken together, these findings suggest that feature design played a more important role than model selection in determining validation performance. In particular, the consistent gains obtained by incorporating time proxy variables, and the superior performance of the FWI + Time feature set across all models, support the value of combining temporally structured proxy information with meteorologically derived wildfire danger indices for wildfire occurrence prediction.

3.2. Predictive Performance on the Test Set

Predictive performance was compared using the independent test set to evaluate out-of-sample generalization. As shown in Figure 6, performance differences among models within the same feature set remained modest, whereas differences across feature sets were more pronounced. This pattern was again most evident for the FWI + Time feature set, where AUC values ranged from 0.8317 to 0.8394, corresponding to a narrow maximum-minimum gap of 0.0077.

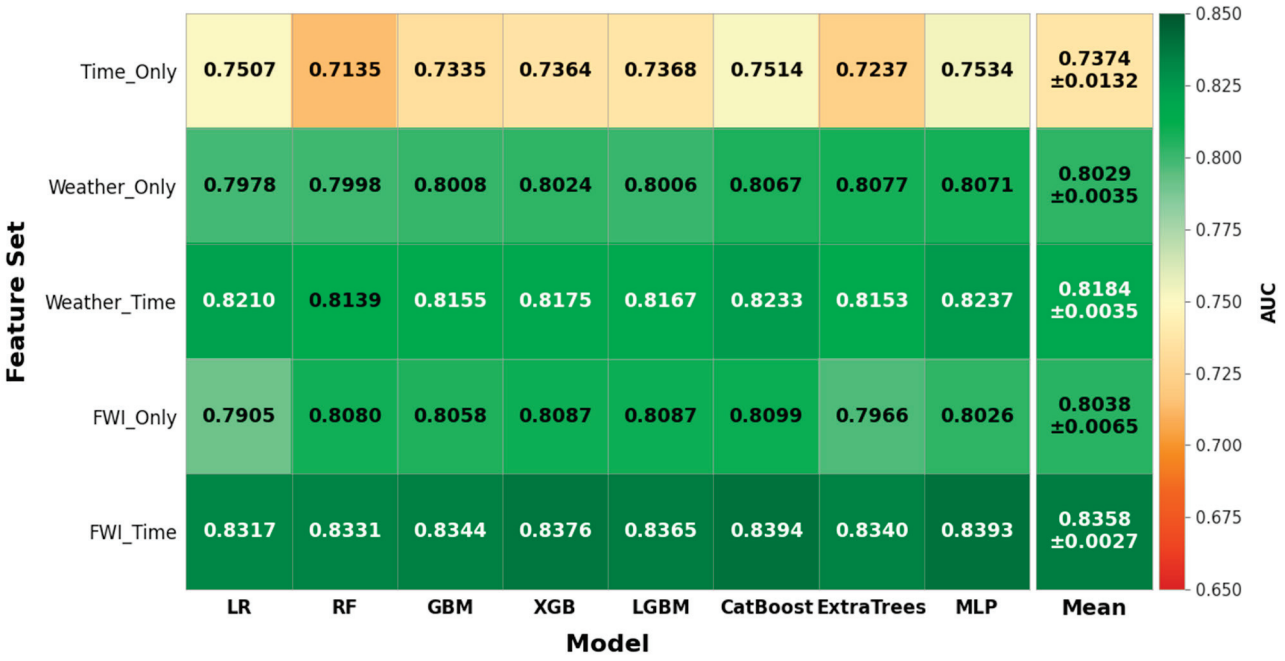


Figure 6. Heatmap of AUC across feature sets and models (Test Set).

Across feature sets, Time-only yielded the lowest performance, with AUC values ranging from 0.7135 to 0.7534 and a mean of 0.7374. Weather-Only and FWI-Only showed similar average performance, with mean AUC values of 0.8029 and 0.8038, respectively, indicating only a marginal advantage for FWI-Only on the test set. By contrast, adding

https://doi.org/10.3390/fire9060217

time proxy variables improved performance in both settings. The mean AUC increased from 0.8029 to 0.8184 for Weather-Only → Weather + Time, and from 0.8038 to 0.8358 for FWI-Only → FWI + Time, with the latter showing the larger gain. These results indicate that time proxy variables contributed consistently to predictive improvement, especially when combined with FWI-based inputs.

Among the models, CatBoost achieved the highest AUC under the FWI + Time feature set (0.8394), followed very closely by MLP (0.8393), XGBoost (0.8376), and LightGBM (0.8365). However, the differences among these top-performing models were negligible, again indicating that no single algorithm was clearly dominant once an informative feature set was used. In contrast, the choice of feature set had a more substantial effect on performance, with FWI + Time consistently outperforming all other feature sets across all models. This supports the conclusion that feature composition, particularly the integration of time proxy variables with FWI, played a more important role than model selection in determining test-set performance.

Figure 7 further illustrates this pattern for the CatBoost model. Figure 7a shows that the ROC curve of FWI + Time lies consistently above those of the other feature sets, confirming its superior discrimination ability. Figure 7b shows that the AUC differences between the validation and test sets remain small across all feature sets: +0.0122 for Time-only, +0.0034 for Weather-Only, +0.0075 for Weather + Time, −0.0030 for FWI-Only, and −0.0017 for FWI + Time. This consistency, combined with the fact that the performance ranking of the feature sets remains identical across both sets, confirms that the model generalizes well. Such a narrow gap suggests that the validation results reliably translate to the test data without significant overfitting.

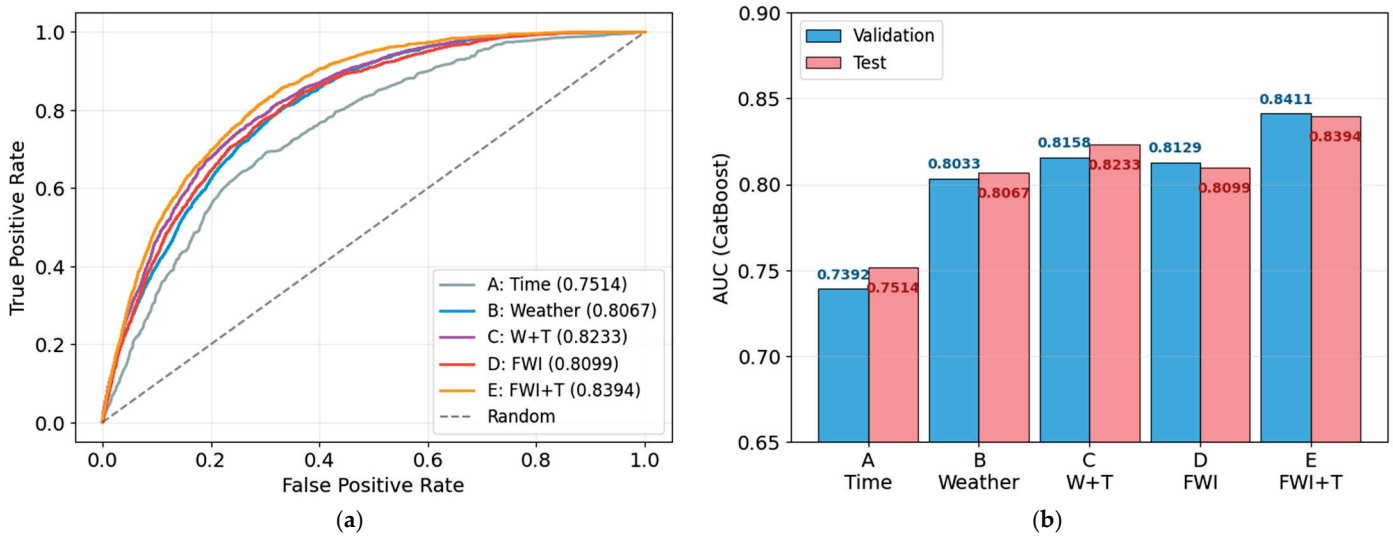


Figure 7. Comparison of performance by feature set (CatBoost). (**a**) ROC curves by feature (Test set). (**b**) AUC: Validation vs. Test.

3.3. Performance Metrics of the Representative Model

Table 12 summarizes the major performance metrics for the best-performing model under each feature set on the test set. Consistent with the AUC-based comparison, the FWI + Time feature set showed the best overall performance, with CatBoost achieving an AUC of 0.8394, a Recall of 0.6019, a Precision of 0.0550, and an F1-score of 0.1008. Among all feature sets, this configuration yielded the highest AUC, Recall, Precision, and F1-score, indicating the most favorable overall balance in detecting wildfire occurrence cases under the given data conditions.

https://doi.org/10.3390/fire9060217

Table 12. Major performance metrics of the best predictive model by feature set (Test Set).

Feature_Set	Model	AUC	Recall	Precision	F1
Time_Only	MLP	0.7534	0.5703	0.0373	0.0700
Weather_Only	Extra Trees	0.8077	0.4349	0.0501	0.0899
Weather_Time	MLP	0.8237	0.5807	0.0535	0.0980
FWI_Only	CatBoost	0.8099	0.5005	0.0528	0.0954
FWI_Time	CatBoost	0.8394	0.6019	0.0550	0.1008

A similar pattern was observed for the intermediate feature sets. Adding time proxy variables improved all major metrics in both meteorological settings: Weather + Time outperformed Weather-Only, and FWI + Time outperformed FWI-Only in terms of AUC, Recall, Precision, and F1-score. This suggests that time proxy variables contributed not only to discrimination performance but also to event detection sensitivity.

Although Precision and F1-score remained relatively low across all feature sets, this is attributable to the highly imbalanced nature of the dataset, in which wildfire occurrence accounted for only a small fraction of all observations. Because evaluation was conducted under the original class distribution to reflect realistic operational conditions, these values should be interpreted as a characteristic of rare-event prediction rather than as a simple limitation of model performance. In this context, the relatively high Recall achieved by FWI + Time supports its practical usefulness for wildfire early warning.

3.4. Comparison Between Machine Learning and Deep Learning Models

Finally, predictive performance was compared between machine learning and deep learning models using the FWI + Time feature set. As shown in Table 13, the differences among models were very small. In terms of AUC, LSTM_Simple showed the highest value (0.8397), followed by CatBoost (0.8394), Deep_MLP (0.8385), and LSTM_Complex (0.8368). The maximum difference across all models was only 0.0029, indicating that overall discrimination performance was highly similar across machine learning and deep learning approaches.

Table 13. Comparison between machine learning and deep learning models (Test Set, FWI + Time).

Type	Model	AUC	Recall	Precision	F1
ML	CatBoost	0.8394	0.6019	0.0550	0.1008
DL	Deep_MLP	0.8385	0.6536	0.0507	0.0941
DL	LSTM_Simple	0.8397	0.6412	0.0531	0.0980
DL	LSTM_Complex	0.8368	0.6468	0.0520	0.0962

In terms of Recall, the deep learning models outperformed CatBoost, whereas CatBoost achieved the highest Precision and F1-score. This suggests a slightly better performance balance for CatBoost despite the higher sensitivity of the deep learning models.

Overall, the differences between machine learning and deep learning models were marginal. This indicates that, under the FWI + Time feature set, competitive predictive performance can be achieved not only with deep learning architectures but also with parsimonious machine learning models.

https://doi.org/10.3390/fire9060217

3.5. Hypothesis Testing Results

3.5.1. RH1: Predictive Power of the Time Proxy Variables

RH1 tests whether a model based solely on time proxy variables (Time-only) performs better than random prediction. On the validation set, the Time-only feature set yielded AUC values ranging from 0.7120 to 0.7495 across models (Table 11), while on the test set the range was 0.7135 to 0.7534 (Figure 6). All values were substantially higher than the random-classification benchmark (AUC = 0.5), indicating that time proxy variables alone retain meaningful predictive power.

These results suggest that calendar-based variables capture recurrent temporal patterns in wildfire occurrence even without explicit meteorological inputs. Therefore, RH1 is supported.

3.5.2. RH2: Performance Improvement of the FWI + Time Model

RH2 examines whether the FWI + Time feature set improves predictive performance relative to FWI-Only. As shown in Table 14, adding time proxy variables improved performance in both meteorological settings on the test set. The mean AUC increased from 0.8029 to 0.8184 for Weather-Only → Weather + Time, corresponding to an improvement of +0.0155 (+1.93%). A larger gain was observed for FWI-Only → FWI + Time, where the mean AUC increased from 0.8038 to 0.8358, corresponding to +0.0320 (+3.98%). In addition, FWI + Time outperformed Weather + Time by +0.0174 (+2.13%).

Table 14. Comparison of mean AUC across feature sets and models (Test Set).

Comparison Feature Set	Mean AUC Difference	Percentage of Improvement [1]
Weather + Time vs. Weather-only	+0.0155	+1.93%
FWI + Time vs. FWI-only	+0.0320	+3.98%
FWI + Time vs. Weather + Time	+0.0174	+2.13%

[1] Note: Performance improvement rate was calculated as: (Compared AUC − Reference AUC)/Reference AUC × 100.

These results indicate that time proxy variables consistently improved predictive performance, with a particularly strong effect when combined with FWI-based inputs. Therefore, RH2 is supported.

3.6. Statistical Validation

To assess whether the observed AUC differences were attributable to chance, bootstrap-based 95% confidence intervals and the DeLong test were applied to the CatBoost results on the test set.

As shown in Table 15 and Figure 8a, FWI + Time achieved the highest AUC (0.8395, 95% CI: [0.8320, 0.8469]), followed by Weather + Time (0.8235, [0.8156, 0.8315]), FWI-Only (0.8098, [0.8012, 0.8175]), Weather-Only (0.8067, [0.7986, 0.8148]), and Time-Only (0.7517, [0.7416, 0.7612]). The confidence interval for FWI + Time did not overlap with that of FWI-Only, indicating that the improvement associated with adding time proxy variables to FWI was unlikely to be due to sampling variability. A similar, though smaller, separation was observed between Weather + Time and Weather-Only.

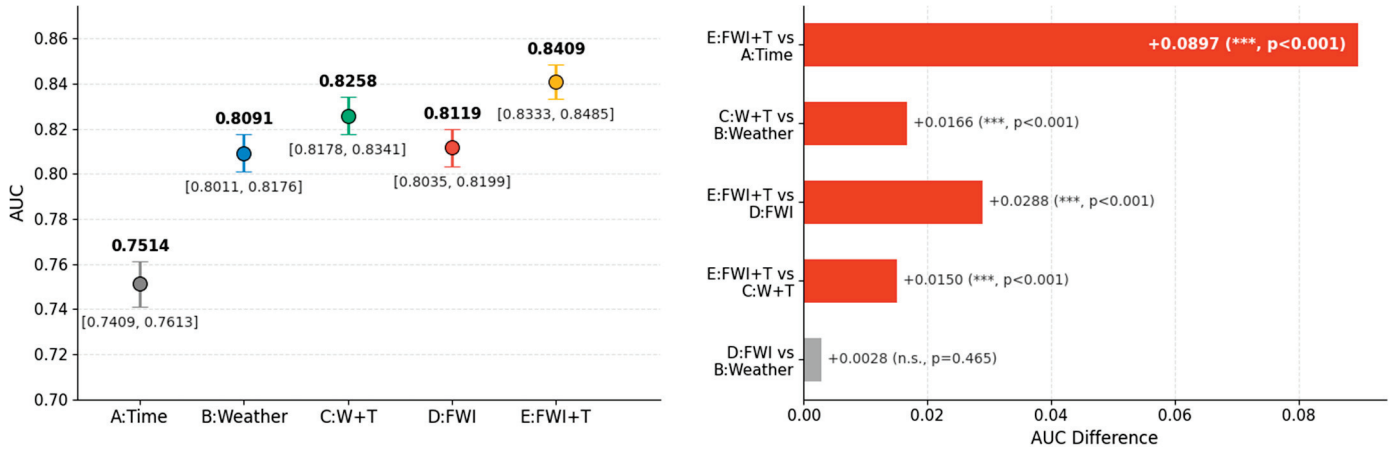
The DeLong test results in Table 16 and Figure 8b further confirmed this pattern. FWI + Time significantly outperformed FWI-Only (ΔAUC = 0.0295, $p < 0.001$), and Weather + Time also improved significantly over Weather-Only (ΔAUC = 0.0166, $p < 0.001$). FWI + Time further showed a statistically significant advantage over Weather + Time (ΔAUC = 0.0160, $p < 0.001$) and a much larger advantage over Time-Only (ΔAUC = 0.0880,

https://doi.org/10.3390/fire9060217

$p < 0.001$). By contrast, the difference between FWI-Only and Weather-Only was small and not statistically significant (ΔAUC = 0.0031, $p = 0.409$).

Table 15. Bootstrap 95% CI (CatBoost, Test Set).

Feature Set	AUC	95% CI (Confidence Interval)
Time-Only	0.7517	[0.7416, 0.7612]
Weather-Only	0.8067	[0.7986, 0.8148]
Weather + Time	0.8235	[0.8156, 0.8315]
FWI-Only	0.8098	[0.8012, 0.8175]
FWI + Time	0.8395	[0.8320, 0.8469]



(**a**) (**b**)

Figure 8. DeLong test & Bootstrap 95% CI (Test Set, CatBoost). (**a**) Bootstrap 95% CI (**b**) DeLong test: AUC Differences.

Table 16. DeLong test results for pairwise AUC comparisons among feature sets (CatBoost, Test Set).

Comparison	AUC1	AUC2	ΔAUC	Z Statistic	*p*-Value	Significant
FWI vs. Weather (same info base)	0.8099	0.8067	0.0031	0.8250	0.409379	n.s.
FWI + Time vs. Weather + Time	0.8394	0.8233	0.0160	6.0079	0.000000	***
FWI + Time vs. FWI-only (Time proxy contribution, FWI)	0.8394	0.8099	0.0295	11.2753	0.000000	***
Weather + Time vs. Weather-only (Time proxy contribution, Weather)	0.8233	0.8067	0.0166	7.5843	0.000000	***
FWI + Time vs. Time-only	0.8394	0.7514	0.0880	22.4678	0.000000	***

Note: AUC1 and AUC2 denote the AUC values of the two compared models. ΔAUC indicates the difference in AUC between the two models (AUC1 − AUC2), and the Z statistic denotes the standardized test statistic from the DeLong test. *** $p < 0.001$, n.s. = not significant ($p \geq 0.05$).

Overall, these results consistently support the superior performance of the FWI + Time model across both confidence interval estimation and paired AUC comparisons. They also show that time proxy variables contribute additional predictive value in both weather-based and FWI-based models, with a larger improvement observed when combined with FWI. Taken together, these findings support RH2 based on multiple statistical evaluation approaches.

https://doi.org/10.3390/fire9060217

3.7. Variable Contribution Analysis

SHAP analysis was used to examine the relative contributions of the input variables to prediction performance on the test set. To reduce model-specific interpretational bias, the analysis was conducted for six tree-based ensemble models (RF, GBM, XGB, LGBM, CatBoost, and Extra Trees). Figure 9 shows the variable-level contribution patterns, with variables ordered by mean importance across models, and Figure 10 summarizes the group-level contribution ratios of the FWI components and time proxy variables.

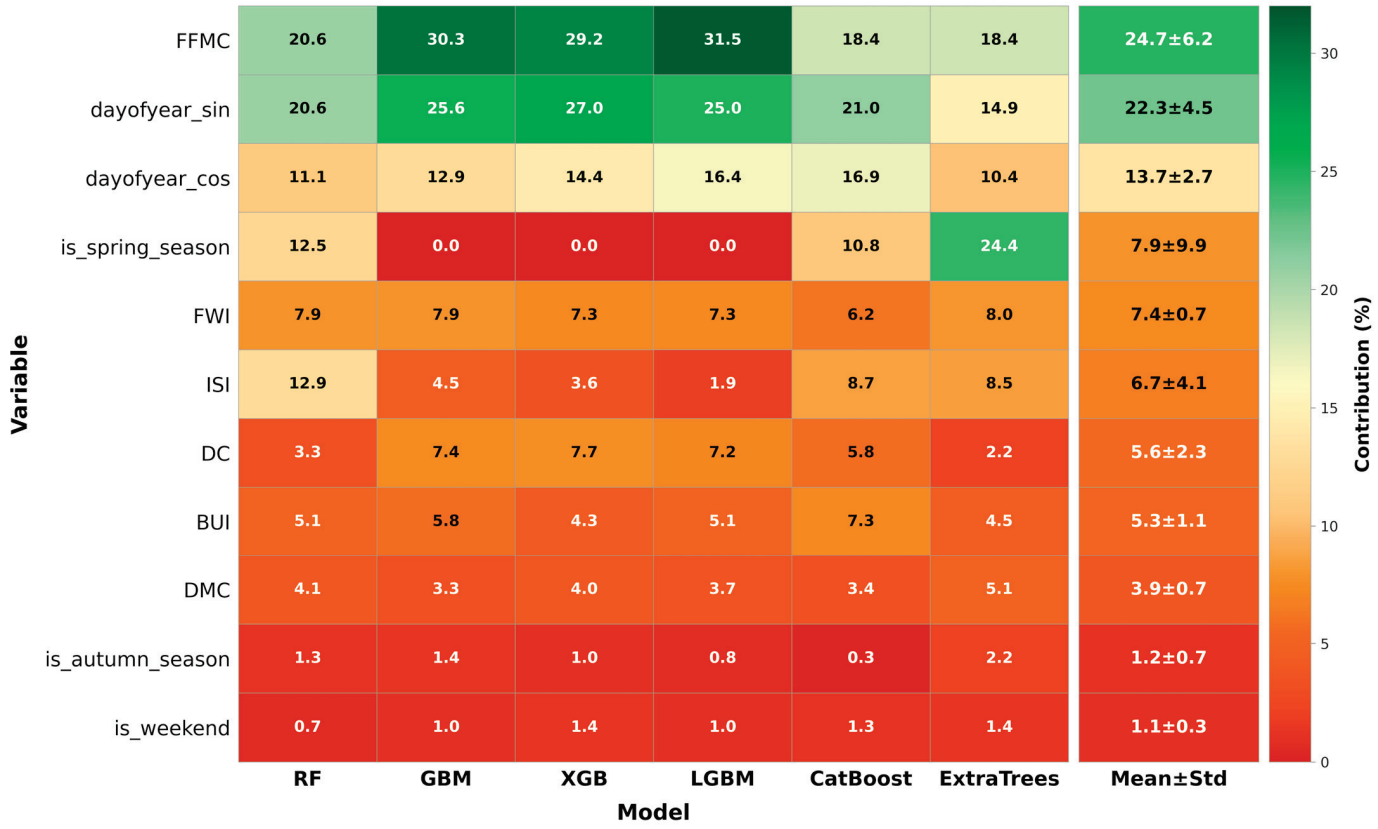


Figure 9. SHAP-based variable importance heatmap by model (%) (Test Set). Note: Variables are ordered by mean contribution across the six tree-based ensemble models. The rightmost column shows the mean ± standard deviation of the SHAP contribution for each variable.

At the variable level (see Figure 9), FFMC showed the highest mean contribution (24.7 ± 6.2%), followed by dayofyear_sin (22.3 ± 4.5%) and dayofyear_cos (13.7 ± 2.7%). This indicates that wildfire occurrence prediction was driven primarily by FFMC, reflecting short-term fine-fuel moisture conditions, and by dayofyear_sin and dayofyear_cos, reflecting annual cyclical timing. Because dayofyear_sin and dayofyear_cos jointly encode annual cyclical timing, the higher SHAP contribution of dayofyear_sin should be interpreted as data- and encoding-specific rather than as evidence that the sine transformation is intrinsically more important than the cosine transformation. By contrast, is_weekend (1.1 ± 0.3%) and is_autumn (1.2 ± 0.7%) contributed only marginally. Among the remaining variables, is_spring_season showed moderate mean importance (7.9 ± 9.9%), but its contribution varied substantially across models, with some models assigning little to no importance while others assigned relatively high importance.

As shown in Figure 10, averaged across the six models, the FWI variables accounted for 53.7% of the total contribution, while the time proxy variables accounted for 46.3%. Although the balance varied somewhat by model, both groups consistently contributed substantial predictive information. In particular, GBM, XGBoost, and LightGBM assigned

https://doi.org/10.3390/fire9060217

relatively greater importance to the FWI group, whereas Extra Trees and CatBoost assigned comparable or greater importance to the time proxy group.

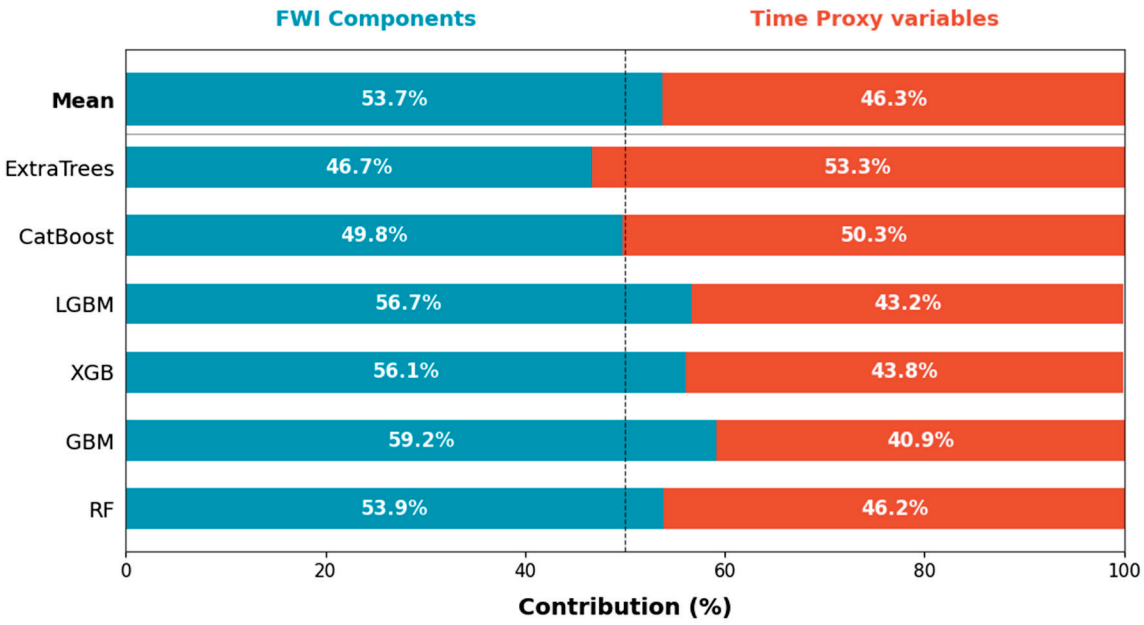


Figure 10. Group-level SHAP contributions by model (%) (Test Set). Note: The stacked bars show the relative contribution of the two variable groups for each model, and the dashed vertical line indicates the 50% reference point.

Taken together, these results suggest that the improved performance of the FWI + Time feature set arose from the complementary effects of two variable groups: FWI captured physically based fire-danger conditions, whereas time proxy variables captured the seasonal concentration structure of wildfire occurrence. This interpretation is consistent with the performance comparisons presented above.

4. Discussion

4.1. Predictive Role of Time Proxy Variables

The results show that time proxy variables provided independent predictive information rather than serving merely as auxiliary inputs. The Time-only feature set yielded AUC values of 0.7120–0.7495 on the validation set and 0.7135–0.7534 on the test set, all of which were well above the random baseline of 0.5. This indicates that wildfire occurrence follows structured temporal regularities and that calendar-based variables can capture recurrent seasonal and behavioral patterns to a meaningful extent.

Their predictive value became more evident when combined with meteorological information. On the test set, the mean AUC increased from 0.8038 for FWI-only to 0.8358 for FWI + Time, corresponding to an improvement of about 4.0%. Moreover, the DeLong test confirmed that this gain was statistically significant ($p < 0.001$), while the difference between FWI-only and Weather-only was not significant. This suggests that the main improvement arose not simply from replacing raw weather variables with FWI, but from adding time proxy information that was not fully captured by either representation alone.

This interpretation is supported by the SHAP analysis. At the group level, the FWI variables accounted for 53.7% of the total contribution, while the time proxy variables accounted for 46.3%. Thus, the improved performance of the FWI + Time feature set appears to reflect a complementary structure in which FWI captures physically based

https://doi.org/10.3390/fire9060217

fire-danger conditions, whereas time proxy variables capture the seasonal concentration structure of wildfire occurrence.

4.2. Comparison with Previous Studies and Methodological Distinctiveness

Recent wildfire prediction studies have increasingly combined meteorological variables with vegetation, topographic, remote-sensing, and socioeconomic indicators to build machine learning models [20,32–34] (Table 17). For example, previous studies have incorporated variables such as the Normalized Difference Vegetation Index (NDVI), biomass, fuel moisture, land use, road accessibility, and population-related indicators to represent fuel conditions and ignition potential [5,7,9,11,16,17,32–34]. Such data-rich approaches are valuable because they can directly or indirectly represent dynamic fuel conditions, vegetation status, landscape characteristics, and human ignition pressure, although they often require diverse data sources and complex integration procedures [7,11,20,32–34].

Table 17. Qualitative methodological comparison between previous studies and the proposed parsimonious framework.

Dimension of Comparison	Previous Studies	Present Study (Parsimonious Framework)
Fuel representation	Vegetation, fuel, environmental, or remote-sensing indicators, such as NDVI, biomass, fuel moisture content, land-cover variables, or fire-danger indices, are often used [7,11,20,32–34].	Indirect representation using components of the Canadian Fire Weather Index (FWI)
Ignition representation	Human-activity or accessibility indicators, such as population density, distance to roads, land use, or socioeconomic variables, are often incorporated [5,9,16,17].	Calendar-based time proxy variables reflecting seasonal fuel dryness and recurrent human activity cycles
Input data sources	Commonly require multi-source integration of meteorological, vegetation, topographic, remote-sensing, or socioeconomic datasets [7,11,20,32–34].	Meteorological observations from Korea Meteorological Administration (KMA) ASOS stations and calendar-based variables
Spatial scale	Varies across studies, including regional, continental, or global-scale modeling with different spatial and temporal resolutions [8,32–34]	Nationwide station-day-based daily prediction across South Korea.
Class imbalance handling	Sampling strategy and class imbalance treatment differ across studies, including resampling or synthetic sampling approaches [20,27,28,32–34]	Preserves the original wildfire occurrence rate in validation and test sets.
Model validation and interpretability	Validation and interpretation strategies differ across studies, including validation datasets, time-series splitting, cross-validation, and variable-importance analysis [20,32–34].	Uses temporal data splitting, bootstrap confidence intervals, DeLong tests, and multi-model SHAP analysis [29–31].

Source: Compiled by the authors based on the literature cited in each row. Note: This table provides a qualitative methodological comparison rather than a direct performance benchmark.

In contrast, the present study adopted a more parsimonious framework that combined the standard Canadian Fire Weather Index (FWI) components with calendar-based time proxy variables. Rather than explicitly incorporating vegetation or socioeconomic datasets, the proposed framework used FWI to represent fuel-related fire danger conditions and time proxies to reflect seasonal fuel dryness and recurrent human activity patterns. Within the evaluation setting of this study, this simpler input structure achieved meaningful predictive

https://doi.org/10.3390/fire9060217

performance for nationwide daily wildfire occurrence prediction (AUC = 0.8394 on the test set).

The methodological distinctiveness of this study lies not only in its parsimonious variable design but also in its evaluation strategy. The original wildfire occurrence rate (1.28%) was preserved in the validation and test sets to better reflect operational conditions. In addition, predictive differences were assessed using bootstrap confidence intervals and the DeLong test [29,30], and the complementary roles of the variable groups were examined through multi-model SHAP analysis including Extra Trees. These procedures strengthen the statistical reliability and interpretability of the proposed framework. Table 17 summarizes the major methodological differences between previous studies and the parsimonious framework proposed here, focusing on qualitative methodological characteristics rather than direct performance benchmarking.

4.3. Operational Implications of Nationwide Wildfire Prediction

The practical value of wildfire prediction should be considered not only at the level of individual locations but also in terms of whether it can provide broader-area early warning signals for operational response. In practice, wildfire suppression is often coordinated at regional and national levels rather than at single observation points. From this perspective, daily nationwide prediction can help translate local prediction outputs into spatially broader risk information that is more relevant to actual response systems.

In this regard, the framework proposed in this study has clear operational advantages. Because it relies only on a small number of meteorological inputs together with calendar-based time proxy variables, it can be applied uniformly across the national observation network without requiring complex integration of vegetation, topographic, or socioeconomic datasets. This parsimonious structure improves scalability and makes the model more suitable for routine nationwide risk monitoring and future linkage with wildfire decision-support systems.

The observed performance level also supports this operational interpretation. The proposed framework achieved an AUC of 0.8394 and a Recall of 0.6019, suggesting that it can function as a useful early warning tool for identifying elevated wildfire risk at the national scale. Although some local false positives and false negatives remain unavoidable, the model appears sufficiently informative to support broader-area preparedness, resource pre-positioning, and coordinated response across regional and central wildfire management systems.

4.4. Limitations and Future Research

Despite these results, several limitations and directions for future research should be noted. First, because this study used daily data, it could not fully capture short-term fire-weather variability at sub-daily scales. Future work should examine whether hourly observations or numerical weather prediction data can improve both predictive performance and early warning capability.

Second, although the time proxy variables used here effectively captured seasonal and recurrent temporal patterns, they remain proxy indicators rather than direct measures of human activity. Future studies may refine this component by incorporating additional calendar-based variables, such as public holidays or farming seasons, while maintaining the parsimonious structure of the framework.

Third, the proposed framework was evaluated using 15 years of nationwide data from South Korea, and its broader generalizability remains to be tested. Additional validation in other regions with strong seasonal wildfire patterns would help clarify the wider applicability of this approach.

https://doi.org/10.3390/fire9060217

Finally, the practical usefulness of the model will depend on how prediction thresholds are set under operational objectives. Future research should therefore explore threshold optimization, cost-sensitive learning, and multi-horizon forecasting beyond same-day prediction in order to strengthen the framework's value as an operational early warning system.

5. Conclusions

This study developed and evaluated a parsimonious framework for daily nationwide wildfire occurrence prediction by combining components of the Canadian Fire Weather Index (FWI) with calendar-based time proxy variables. The main objective was to examine whether time proxy variables could provide meaningful predictive information beyond meteorological inputs and thereby improve operationally useful wildfire early warning at the national scale.

The results showed that time proxy variables were not merely auxiliary inputs. Even the Time-only feature set performed clearly above the theoretical random-classifier baseline of AUC = 0.5, indicating that wildfire occurrence in South Korea follows structured temporal regularities. More importantly, the FWI + Time feature set consistently outperformed FWI-only, with the mean test AUC increasing from 0.8038 to 0.8358, and this improvement was supported by the DeLong test ($p < 0.001$). SHAP analysis further showed that the FWI variables accounted for 53.7% of the total contribution and the time proxy variables for 46.3%, indicating that the performance gain arose from a complementary structure rather than simple redundancy.

From an operational perspective, the proposed framework achieved AUC = 0.8394 and Recall = 0.6019, suggesting that it can provide useful broader-area wildfire risk signals for resource pre-positioning and coordinated regional or national response. Because the model relies only on a limited number of meteorological variables and calendar-based time proxy variables, it offers a practical and scalable alternative for nationwide wildfire early warning without requiring complex ancillary datasets.

Overall, this paper demonstrates that meaningful nationwide wildfire occurrence prediction can be achieved using a relatively simple input structure. Future research should further examine the generalizability of this approach in other regions with strong seasonal wildfire patterns and improve its operational utility through threshold optimization, cost-sensitive learning, and multi-horizon forecasting.

Author Contributions: Conceptualization, B.C.; methodology, B.C.; software, B.C.; validation, B.C.; formal analysis, B.C.; investigation, B.C.; data curation, B.C.; writing—original draft preparation, B.C.; writing—review and editing, B.C. and G.-Y.K.; visualization, B.C.; supervision, G.-Y.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Institute of Information and Communications Technology Planning and Evaluation (IITP)-Innovative Human Resource Development for Local Intellectualization Program grant funded by the Korean Government [Ministry of Science and ICT (MSIT)] under Grant IITP-2026-RS-2022-00156360.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used in this study were obtained from the Korea Forest Service wildfire statistics and the Korea Meteorological Administration Open MET Data Portal. These data are publicly available from the respective official sources, although some pre-processing and integration were performed by the author for the purposes of this study. The processed data are available from the corresponding author on reasonable request.

https://doi.org/10.3390/fire9060217

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

ASOS	Automated Synoptic Observing System
AUC	Area under the receiver operating characteristic curve
Deep_MLP	Deep Multi-Layer Perceptron
FWI	Fire Weather Index
LR	Logistic Regression
LSTM	Long Short-Term Memory
LSTM_Complex	Long Short-Term Memory with stacked layers or higher complexity
LSTM_Simple	Long Short-Term Memory with simplified architecture
MLP	Multi-Layer Perceptron (Deep Learning model)
RF	Random Forest
RH	Research Hypothesis
SHAP	SHapley Additive exPlanations

References

1. IPCC. *Climate Change 2022: Impacts, Adaptation, and Vulnerability*; Cambridge University Press: Cambridge, UK, 2022.
2. Jones, M.W.; Canadell, J.G.; Jackson, R.B.; Peters, G.P.; Rothe, M.; van Vuuren, D.P. Global and regional trends and drivers of fire under climate change. *Rev. Geophys.* **2022**, *60*, e2020RG000726. [CrossRef]
3. Senande-Rivera, M.; Campos-Rodríguez, R.; Rodrigues, M.; Russo, A.; Toumi, R. Global wildfire patterns and drivers under climate change. *Biogeosciences* **2025**, *22*, 7591–7610. [CrossRef]
4. Van Wagner, C.E. *Development and Structure of the Canadian Forest Fire Weather Index System*; Forestry Technical Report 35; Canadian Forestry Service: Ottawa, ON, Canada, 1987.
5. Prestemon, J.P.; Hawbaker, T.J.; Bowden, J.H.; Carpenter, R.J.; Abt, K.L.; Noonan, C.F.; Stapleton, E.M.; Adams, M.D. *Wildfire Ignitions: A Review of the Science and Recommendations for Empirical Modeling*; General Technical Report SRS-171; USDA Forest Service: Asheville, NC, USA, 2013.
6. Plucinski, M.P. *A Review of Wildfire Occurrence Research; Fire Note 94*; Bushfire CRC: Melbourne, Australia, 2012.
7. Alipour, M.; La Puma, I.; Picotte, J.; Shamsaei, K.; Rowell, E.; Watts, A.; Kosovic, B.; Ebrahimian, H.; Taciroglu, E. A multimodal data fusion and deep learning framework for large-scale wildfire surface fuel mapping. *Fire* **2023**, *6*, 36. [CrossRef]
8. Di Giuseppe, F.; Barnard, C.; Lopez-Gomez, I.; Drusch, M.; Jackson, T.; Kim, H.; Jung, M.; Kerr, Y.; Leroux, D.; Pan, M.; et al. Global data-driven prediction of fire activity. *Nat. Commun.* **2025**, *16*, 58097. [CrossRef]
9. Costafreda-Aumedes, S.; Comas, C.; Vega-García, C. Human-Caused Fire Occurrence Modelling in Perspective: A Review. *Int. J. Wildl. Fire* **2017**, *26*, 983–998. [CrossRef]
10. Moritz, M.A.; Parisien, M.-A.; Batllori, E.; Krawchuk, M.A.; Van Dorn, J.; Ganz, D.J.; Hayhoe, K. Climate change and disruptions to postfire recovery. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, E2858.
11. Chuvieco, E.; Yebra, M.; Martino, S.; Thonicke, K.; Gómez-Giménez, M.; San-Miguel, J.; Oom, D.; Velea, R.; Mouillot, F.; Molina, J.R.; et al. Towards an Integrated Approach to Wildfire Risk Assessment: When, Where, What and How May the Land scapes Burn. *Fire* **2023**, *6*, 215. [CrossRef]
12. Moritz, M.A.; Parisien, M.-A.; Batllori, E.; Krawchuk, M.A.; Van Dorn, J.; Ganz, D.J.; Hayhoe, K. Climate Change and Disruptions to Global Fire Activity. *Ecosphere* **2012**, *3*, 49. [CrossRef]
13. De Groot, W.J. Interpreting the Canadian forest fire weather index system. In *Proceedings of the Fourth Central Regional Fire Weather Committee Scientific and Technical Seminar*; Canadian Forestry Service: Ottawa, ON, Canada, 1987; pp. 3–14.
14. Bedia, J.; Golding, N.; Casanueva, A.; Iturbide, M.; Herrera, S.; Gutiérrez, J.M. Seasonal predictions of fire weather index: Paving the way for their operational applicability in Mediterranean Europe. *Clim. Serv.* **2018**, *9*, 101–110. [CrossRef]
15. de Jong, M.C.; Wooster, M.J.; Kitchen, K.; Manley, C.; Gazzard, R.; McCall, F.F. Calibration and Evaluation of the Canadian Forest Fire Weather Index (FWI) System for Improved Wildland Fire Danger Rating in the United Kingdom. *Nat. Hazards Earth Syst. Sci.* **2016**, *16*, 1217–1237. [CrossRef]
16. Balch, J.K.; Bradley, B.A.; Abatzoglou, J.T.; Nagy, R.C.; Fusco, E.J.; Mahood, A.L. Human-started wildfires expand the fire niche across the United States. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 2946–2951. [CrossRef]

https://doi.org/10.3390/fire9060217

17. Pereira, J.M.C.; Turkman, M.A.A.; Turkman, K.F.; Oom, D. Anthromes displaying evidence of weekly cycles in active fire data cover 70% of the global land surface. *Sci. Rep.* **2019**, *9*, 11424. [CrossRef]
18. Park, I.W.; Mann, M.L.; Flint, L.E.; Flint, A.L.; Moritz, M.A. Relationships of climate, human activity, fire history, and pyrome to spatiotemporal variation in annual fire probability across California. *PLoS ONE* **2021**, *16*, e0254723. [CrossRef]
19. Bernardi, M.; Petrella, L. Multiple Seasonal Cycles Forecasting Model: The Italian Electricity Demand. *Stat. Methods Appl.* **2015**, *24*, 671–695. [CrossRef]
20. Jain, P.; Coogan, S.C.P.; Subramanian, S.G.; Crowley, M.; Taylor, S.; Flannigan, M. A review of machine learning applications in wildfire science and management. *Environ. Rev.* **2020**, *28*, 478–505. [CrossRef]
21. Van Wagner, C.E.; Pickett, T.L. *Equations and FORTRAN Program for the Canadian Forest Fire Weather Index System*; Forestry Technical Report 33; Canadian Forestry Service: Ottawa, ON, Canada, 1985.
22. Quilcaillé, Y.; Gudmundsson, L.; Seneviratne, S.I. FWI_CMIP6: Fire Weather Index Calculations for CMIP6 Models. GitHub Repository, 2023. Available online: https://github.com/yquilcaille/FWI_CMIP6 (accessed on 21 March 2026).
23. Wang, X.; Wotton, B.M.; Cantin, A.S.; Parisien, M.-A.; Anderson, K.; Moore, B.; Flannigan, M.D. cffdrs: An R package for the Canadian Forest Fire Danger Rating System. *Ecol. Process.* **2017**, *6*, 5. [CrossRef]
24. Fisher, N.I. *Statistical Analysis of Circular Data*; Cambridge University Press: Cambridge, UK, 1993.
25. Rasmussen, C.E.; Williams, C.K.I. *Gaussian Processes for Machine Learning*; MIT Press: Cambridge, MA, USA, 2006.
26. Bentley, J.L. Multidimensional Binary Search Trees Used for Associative Searching. *Commun. ACM* **1975**, *18*, 509–517. [CrossRef]
27. Saito, T.; Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE* **2015**, *10*, e0118432. [CrossRef]
28. He, H.; Garcia, E.A. Learning from Imbalanced Data. *IEEE Trans. Knowl. Data Eng.* **2009**, *21*, 1263–1284. [CrossRef]
29. Efron, B.; Tibshirani, R.J. *An Introduction to the Bootstrap*; Chapman & Hall/CRC: New York, NY, USA, 1993.
30. DeLong, E.R.; DeLong, D.M.; Clarke-Pearson, D.L. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics* **1988**, *44*, 837–845. [CrossRef]
31. Lundberg, S.M.; Lee, S.-I. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30*; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2017; pp. 4765–4774.
32. Liu, Y.; Huang, H.; Wang, S.-C.; Zhang, T.; Xu, D.; Chen, Y. ELM2.1-XGBFire1.0: Improving wildfire prediction by integrating a machine learning fire model in a land surface model. *Geosci. Model Dev.* **2025**, *18*, 4103–4117. [CrossRef]
33. Kondylatos, S.; Prapas, I.; Ronco, M.; Papoutsis, I.; Camps-Valls, G.; Piles, M.; Fernandez-Torres, M.A.; Carvalhais, N. Wildfire danger prediction and understanding with deep learning. *Geophys. Res. Lett.* **2022**, *49*, e2022GL099368. [CrossRef]
34. Lee, C.; Choi, E.H.; Han, Y.; Lee, Y. Year-round daily wildfire prediction and key factor analysis using machine learning: A case study of Gangwon State, South Korea. *Sci. Rep.* **2025**, *15*, 15508. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

https://doi.org/10.3390/fire9060217

fire

MDPI

Article

Research on Gas Concentration Prediction Method Based on Decoupling of Temporal Feature and Dynamic Relationship Reconstruction

Yongle Yan [1,*], Yichao Zhao [2] and Jiuwu Hui [3]

[1] School of Management, Xi'an University of Science and Technology, Xi'an 710054, China
[2] College of Energy and Mining Engineering, Xi'an University of Science and Technology, Xi'an 710054, China
[3] Department of Building Environment and Energy Engineering, The Hong Kong Polytechnic University, Hong Kong SAR, China
* Correspondence: 13191490924@163.com

Abstract

Accurate multi-channel gas concentration prediction is very important for coal mine safety. However, the dynamic reconstruction of the sensor network often interferes with the input sequence. Existing models face a critical trade-off: channel-independent models are robust to sequence changes but ignore spatial coupling, while channel-dependent models overfit fixed sequences, leading to performance collapse during rearrangements. This paper presents a gas concentration prediction framework based on channel permutation-invariant interaction (CPiRi) to reconcile these limitations. CPiRi employs a spatio-temporal decoupling architecture where a frozen univariate pre-trained encoder independently extracts temporal features to ensure sequence robustness. Subsequently, a permutation-equivariant spatial module utilizes self-attention to model inter-channel gas emission relationships based on data content rather than positional indices. To achieve true permutation invariance, we introduce channel-shuffling regularization during training, forcing the model to learn content-driven relational reasoning. Evaluations on 15 real-world Chinese coal mine datasets demonstrate that CPiRi achieves highly competitive accuracy and consistently outperforms mainstream baselines in both prediction precision and structural adaptability. This study offers a robust technical pathway for gas monitoring in dynamic environments, substantially improving the reliability of intelligent mine safety systems.

Keywords: gas concentration prediction; multi-variable time series; channel permutation invariance; spatio-temporal decoupling; deep learning

1. Introduction

Under conditions of high-intensity and continuous underground mining, gas emission exhibits pronounced temporal variability and spatial heterogeneity. Localized, transient accumulations can readily exceed safety thresholds and trigger cascading disasters, constituting one of the primary causes of major and catastrophic coal mine accidents. Compared with post-incident mitigation, proactive time-series prediction of gas concentration based on continuous multi-point monitoring data directly determines the lead time of early warnings and the effective compression of emergency response windows. It has therefore become a critical component in intelligent mine safety management systems [1].

With the large-scale upgrading of underground monitoring infrastructures, dense deployment of multi-sensor networks has been realized across entire mines. Key locations—

including mining faces, return airways, and critical nodes—are now equipped with stable multi-channel synchronous acquisition systems, forming a comprehensive multivariate time-series data foundation [2]. This engineering advancement provides practical data support for cross-sensor coupled modeling and high-precision prediction, thereby enabling gas concentration forecasting methods that capture global collaborative evolution patterns to be feasibly implemented in real-world scenarios. However, the actual underground operational environment is far from an idealized data acquisition setting. Influenced by harsh conditions such as high humidity, heavy dust, and strong electromagnetic interference, gas sensors require periodic offline maintenance and calibration, with frequent hardware replacement. Meanwhile, routine maintenance and local adjustments of mine-monitoring network systems are often accompanied by sensor replacement, re-numbering, and channel reordering, resulting in structural inconsistencies in the input data indexing. Under these routine operational processes, sensor identification systems are inevitably reset, while the number and ordering of data channels in transmission links continuously vary, leading to globally non-stationary and structurally inconsistent input data. More critically, most existing time-series prediction models implicitly assume a one-to-one correspondence between channel indices and physical sensing points. During training, model parameters become tightly coupled with fixed channel positions. Once channel reordering or partial replacement occurs in practice, the learned inter-sensor relationships are disrupted, causing systematic prediction bias or even complete model failure. This issue severely constrains the sustained reliability of online early warning systems in dynamic operational environments.

From the perspective of existing research, multichannel gas time-series prediction methods generally evolve along two major technical paradigms, both of which struggle to meet the demands of complex underground conditions. The first is the channel-independent modeling paradigm, where each sensing point is treated as an individual univariate time series and modeled separately. Such methods are inherently invariant to channel permutations and maintain basic functionality after sensor replacement or channel reconfiguration [3]. However, they do not explicitly account for the spatial coupling effects of gas migration among adjacent working faces and within the same mining area, failing to capture inter-sensor interaction patterns. As a result, their predictive response is often delayed during abnormal gas fluctuations, with limited accuracy ceilings. The second paradigm is multichannel joint modeling, which leverages multivariate inputs to extract spatiotemporal collaborative features, achieving higher predictive accuracy under stable channel configurations. Nevertheless, these methods rely on fixed-channel ordering and indexing during training, leading to strong positional dependencies in model parameters. Even minor perturbations in input structure can cause significant performance degradation, making them unsuitable for dynamically evolving underground environments. In summary, existing approaches remain fundamentally divided between robustness to channel perturbation and capability for spatial coupling modeling, lacking a unified framework that effectively reconciles both requirements [4].

To address the aforementioned engineering pain points and methodological limitations, this paper proposes a gas concentration prediction framework for underground dynamic sensing scenarios, namely CPiRi, which achieves channel-permutation-invariant modeling through temporal feature disentanglement and dynamic relationship reconstruction. First, a spatiotemporally disentangled architecture is established at the model front end. A univariate time-series encoder is pretrained on historical data from the entire mine and its parameters are subsequently frozen during downstream training, enabling each channel to independently extract both long- and short-term evolutionary features of the corresponding sensing point, thereby inheriting the robustness of independent modeling to channel permutations at a mechanistic level. Second, in the feature interaction stage, a

https://doi.org/10.3390/fire9070267

permutation-equivariant spatial perception module is introduced. Rather than relying on fixed index positions, it adaptively learns the interaction strengths among sensing points according to the real-time data distribution, and dynamically reconstructs the spatial dependency structure of gas emission through a lightweight attention mechanism, so that the model can maintain stable relational representation capability even under channel-order variations. Furthermore, a channel random shuffling strategy is incorporated into the training process, whereby the input sequences are randomly permuted multiple times to explicitly break positional priors and force the model to learn from the intrinsic statistical regularities of the data, thereby acquiring generalization capability with respect to arbitrary channel arrangements. Through the coordinated design of these components, the proposed framework maintains stable predictive performance under practical operational scenarios heavily affected by channel reordering and sensor index resetting.

The core contributions of this paper are pragmatically concentrated in three aspects. First, targeting the long-neglected complex real-world scenario of dynamic sensing operations in coal mines, this study introduce the core mechanism of channel-permutation invariance into the field of multichannel underground gas concentration time-series prediction, thereby breaking the long-standing industrial dilemma in which robustness and predictive accuracy are often treated as mutually incompatible. In doing so, it effectively addresses the adaptability shortcomings of intelligent safety monitoring systems in dynamic operational scenarios. Second, at the methodological level, a lightweight permutation-equivariant spatial interaction coupling module is innovatively designed, together with a channel-shuffling regularization strategy specifically tailored for industrial time-series scenarios, enabling a fully data-driven dynamic reconstruction of cross-sensor gas spatial relationships while effectively reducing the reliance of the model on human-imposed priors, such as hardware identifiers and fixed channel positions. Third, based on full-scale measured gas datasets collected from multiple real underground working faces across the country, comprehensive validation is conducted. The results demonstrate that the proposed method simultaneously offers high accuracy, strong robustness, and low deployment cost. It can be directly integrated into existing intelligent mine monitoring systems for rapid engineering implementation, thereby providing a practical and reliable new technical pathway for precise and advanced gas warning in dynamic high-risk mining scenarios.

2. Literature Review

During underground mining operations, gas exhibits pronounced spatiotemporal heterogeneity across key regions such as working faces, return airways, and corner zones, making continuous multi-point monitoring a fundamental prerequisite for safe production. Early studies on gas concentration prediction primarily relied on statistical methods, fitting historical data under stationarity assumptions, such as short-term trend extrapolation based on autoregressive moving-average models. Under relatively stable ventilation conditions and weak mining disturbances, these methods could provide basic predictive capability [5]. With the accumulation of monitoring data and improvements in computational resources, some studies introduced machine learning approaches, including support vector machines and shallow neural networks, to enhance the characterization of nonlinear fluctuations through nonlinear mapping, achieving moderate accuracy improvements within localized time windows [6]. However, from the perspective of real underground conditions, gas emission is governed by multiple coupled factors, including mining advancement, blasting operations, and stress redistribution in surrounding rock. The resulting process often exhibits strong non-stationarity, characterized by alternating bursts of rapid intensification and decay [7]. When confronted with such abrupt disturbances, the aforementioned methods generally suffer from delayed response and error accumulation in long-horizon

https://doi.org/10.3390/fire9070267

prediction, making them inadequate for the stringent requirements of accuracy and timeliness in high-risk early-warning scenarios [8]. As mine sensing networks have evolved from single-point deployment to full-scale multi-point collaborative monitoring, research has gradually shifted from univariate short-term prediction toward multivariate spatiotemporal coupled modeling, aiming to capture the overall evolution of gas migration by integrating information from different locations. This transition has laid the data and problem foundations for subsequent developments in multichannel prediction methods [9].

Within the multichannel modeling paradigm, channel-independent time-series modeling represents one of the earliest solutions to achieve practical engineering deployment and has long dominated real-world systems [10]. Such methods treat each sensing point as an independent modeling unit, focusing on feature extraction from one-dimensional time series [11]. Representative techniques include temporal convolution for capturing local patterns, lightweight recurrent structures for modeling historical dependencies, and single-channel attention mechanisms to enhance feature representation at critical time intervals [12]. Due to the independent processing across channels, these models do not rely on a unified multichannel input structure during training or inference [13,14]. Consequently, when underground sensors experience temporary failures, device replacement, or the addition of new sensing points, the system can continue operating without reconstructing the overall model, demonstrating strong engineering adaptability and operational convenience. This characteristic has led to their widespread adoption in early intelligent monitoring systems [15,16]. However, from the perspective of gas occurrence and migration mechanisms, significant spatial coupling exists among sensing points within the same mining area. Mining-induced disturbances often propagate regionally, generating coordinated responses across multiple monitoring locations. Channel-independent modeling structurally disconnects such cross-sensor relationships, relying solely on historical data from individual points for extrapolation [17]. As a result, the detection of precursor signals for abnormal gas emissions depends on single-point variations, making it difficult to capture multi-point coordinated changes in a timely manner. During phases of rapid high-concentration accumulation, these methods often exhibit prediction lag or amplitude underestimation, failing to meet the strict accuracy requirements of high-level safety warnings. Their performance limitations are thus fundamentally constrained by the underlying modeling mechanism.

To overcome the limitations of insufficient utilization of single-point information, researchers have progressively shifted toward global multichannel joint modeling, which simultaneously processes time series from multiple sensing points within a unified framework to uncover intrinsic spatiotemporal coupling relationships [18]. Representative approaches include spatiotemporal convolutional structures for modeling local correlations among adjacent sensing points, multichannel joint attention mechanisms for adaptively weighting critical channels and key time steps, and graph-based spatiotemporal inference models designed to capture cross-regional information propagation under complex topologies [19]. These methods are theoretically more consistent with the regional migration characteristics of underground gas and have demonstrated higher predictive accuracy and better multi-point synchronization in practical applications. In particular, they show greater sensitivity in capturing concentration trends during periods of significant mining disturbance. However, such approaches typically assume a stable one-to-one correspondence between input channels and physical sensing points. During training, spatial dependencies are learned based on fixed input ordering, embedding implicit positional coupling information within model parameters. Once sensor replacement, re-indexing, or channel reordering occurs during underground operations, the original input structure is disrupted, and the spatial topology mapping on which the model depends becomes invalid, leading

https://doi.org/10.3390/fire9070267

to significant deviations in inference results [20]. In real production systems, this strong dependence on channel ordering prevents long-term stable operation, often necessitating model retraining after each structural adjustment. This not only increases maintenance costs but also limits applicability in dynamically evolving environments. A comprehensive examination of these two mainstream approaches reveals a clear trade-off between engineering adaptability and predictive accuracy in current multichannel gas concentration prediction technologies. Channel-independent modeling offers strong robustness to channel perturbations but sacrifices accuracy due to neglect of spatial coupling; global joint modeling effectively leverages multichannel information to improve performance but is highly sensitive to fixed channel structures and thus struggles to adapt to dynamic changes in underground sensing systems. Figure 1 illustrates the main research progression in this field.

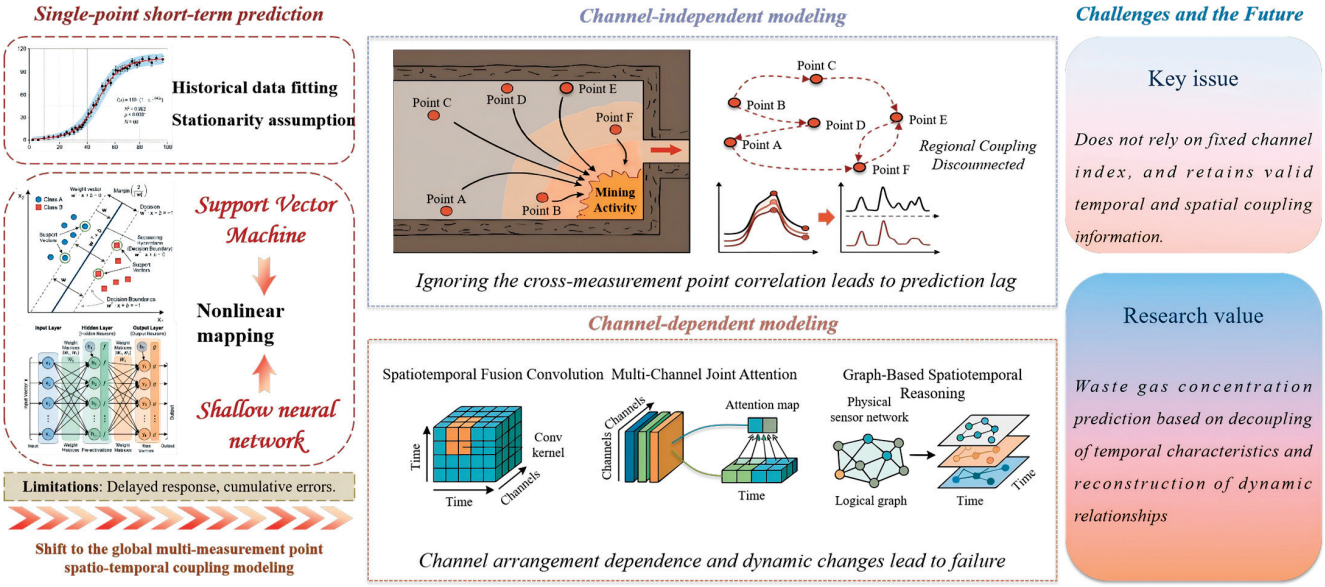


Figure 1. Main research process of gas concentration prediction methods.

Further analysis indicates that most existing studies are developed under the assumption of static sensing networks, lacking dedicated designs for practical operational scenarios such as sensor re-indexing and dynamic channel addition or removal. At the feature level, temporal dependencies and spatial correlations are often entangled in unified modeling processes, with limited explicit disentanglement mechanisms, thereby restricting model generalization under structural variations. In terms of relationship modeling, most approaches rely on predefined topologies or fixed adjacency structures, and data-driven dynamic reconstruction of spatial relationships remains underexplored. Therefore, a critical challenge in this field is how to retain effective spatiotemporal coupling information across multiple sensing points without relying on fixed channel indices, while achieving robust adaptation to channel permutations and structural perturbations. Addressing this interdisciplinary bottleneck between engineering practice and algorithm design, research on gas concentration prediction based on temporal feature disentanglement and dynamic relationship reconstruction presents clear practical necessity and significant technical value.

https://doi.org/10.3390/fire9070267

3. Materials and Methods

3.1. Problem Definition

Consider a multivariate time series consisting of M variables. The historical observations within a time window of length W can be represented as:

$$X = \{x_1, x_2, \ldots, x_W\} \in \mathbb{R}^{W \times M} \tag{1}$$

where W denotes the length of the historical sequence, and $x_t \in \mathbb{R}^M$ represents the observations of all variables at time step t.

The objective of multivariate time-series forecasting is to predict the values over the next H future time steps based on the historical sequence:

$$Y = \{x_{W+1}, \ldots, x_{W+H}\} \in \mathbb{R}^{H \times M} \tag{2}$$

Accordingly, the task can be formulated as learning a mapping function:

$$f_\Theta : \mathbb{R}^{W \times M} \rightarrow \mathbb{R}^{H \times M} \tag{3}$$

such that:

$$\hat{Y} = f_\Theta(X) \tag{4}$$

where θ denotes the model parameters, and $\hat{Y}$ represents the predicted results.

Multivariate time-series forecasting requires not only capturing the dynamic temporal evolution of individual variables, but also modeling the interactions among different variables, i.e., jointly considering temporal dependency and inter-variable interaction. However, existing studies generally follow two distinct paradigms.

One line of work adopts a channel-independent strategy, where each variable is modeled separately. This approach improves model stability and generalization to some extent, but often ignores the latent correlation structure among variables, making it difficult to fully exploit multivariate information. The other line of work focuses on cChannel-dependent modeling, which explicitly captures dependencies across variables to learn inter-variable interaction patterns. Nevertheless, such models tend to rely on a fixed variable ordering during training, leading to performance degradation when variables are permuted or when the variable set changes [21].

To simultaneously account for temporal modeling capability and inter-variable relationship learning, this paper proposes a novel forecasting framework, termed CPiRi, for the first time. A comparison between the proposed framework and conventional modeling strategies is illustrated in Figure 2. By constructing a permutation-invariant relational modeling mechanism, CPiRi preserves strong temporal feature representation while enabling robust learning of inter-variable interactions, thereby enhancing both the stability and generalization performance of multivariate time-series forecasting.

3.2. Overview of the CPiRi Framework

To model temporal patterns and inter-variable relationships simultaneously, the proposed CPiRi framework adopts a temporal–relational decoupling paradigm, which structurally isolates temporal feature learning from variable interaction modeling. The overall architecture of CPiRi is illustrated in Figure 3.

The model primarily comprises three stages:

Stage 1: Temporal feature encoding independently extracts temporal features for the historical sequence of each variable to obtain variable-level temporal representations.

https://doi.org/10.3390/fire9070267

Stage 2: Permutation-invariant relational interaction learns the interaction relationships between variables through a relational modeling module while ensuring the model remains robust to changes in variable ordering.

Stage 3: Independent forecast generation generates future sequence predictions based on the updated variable representations.

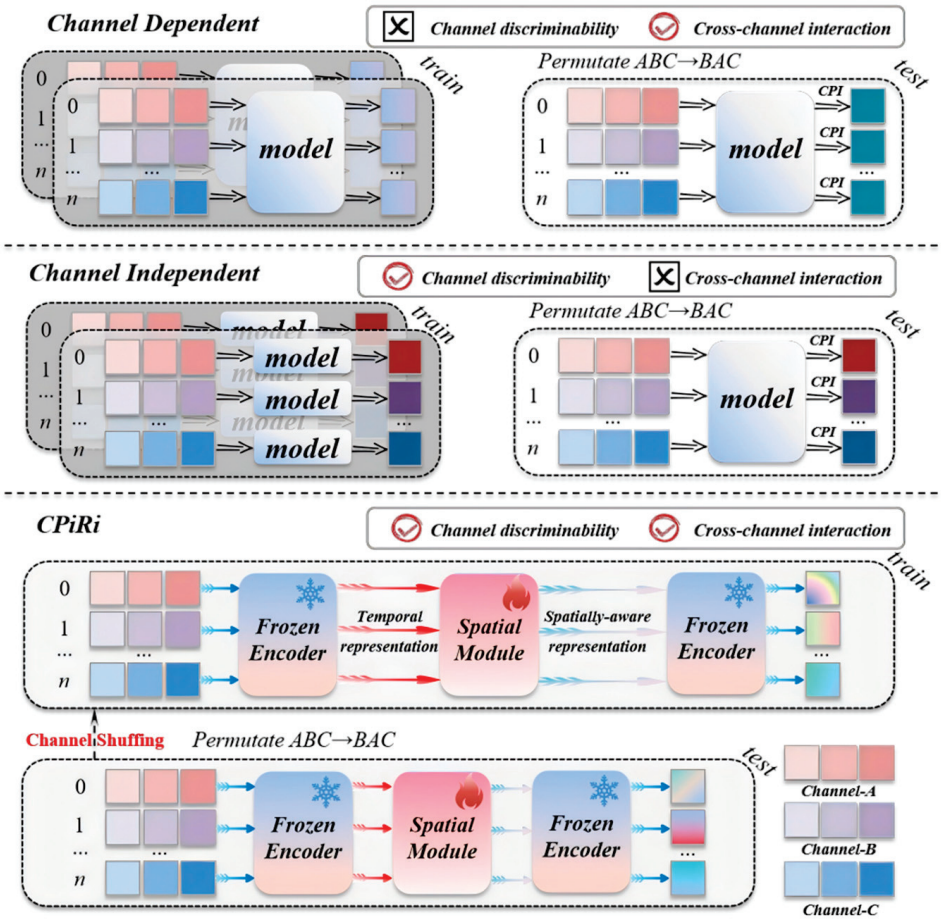


Figure 2. Comparison Chart of Traditional Modeling Strategies and the CPiRi Framework.

The CPiRi framework is characterized by two pivotal features. Traditional multivariate time-series forecasting models typically learn temporal dynamics and inter-variable interactions within a monolithic network structure [22]. Such coupled modeling often increases model complexity and may lead to overfitting on specific data structures. To address this, CPiRi structurally separates temporal modeling from relational modeling: the temporal encoding module focuses on extracting the dynamic patterns of individual variables, while the relational modeling module specifically captures the interaction structures between variables. This decoupled design enables the model to effectively capture inter-variable dependencies while maintaining strong temporal representation capabilities, thereby enhancing overall stability and generalization.

In practical applications, the sequence of variables in multivariate data is often not fixed; for instance, sensor renumbering, the addition of new variables, or the absence of certain variables can lead to permutations in the input order. Models that rely heavily on a fixed variable sequence are prone to performance degradation under these conditions. Consequently, CPiRi introduces a permutation-invariant learning strategy during the training process. By employing techniques such as random variable permutation, the relational modeling module is trained to learn semantic relationships based on variable

https://doi.org/10.3390/fire9070267

content rather than positional indices. This strategy effectively mitigates the model's dependency on variable ordering, ensuring stable predictive performance even when faced with variable reordering or dynamic changes in the variable set.

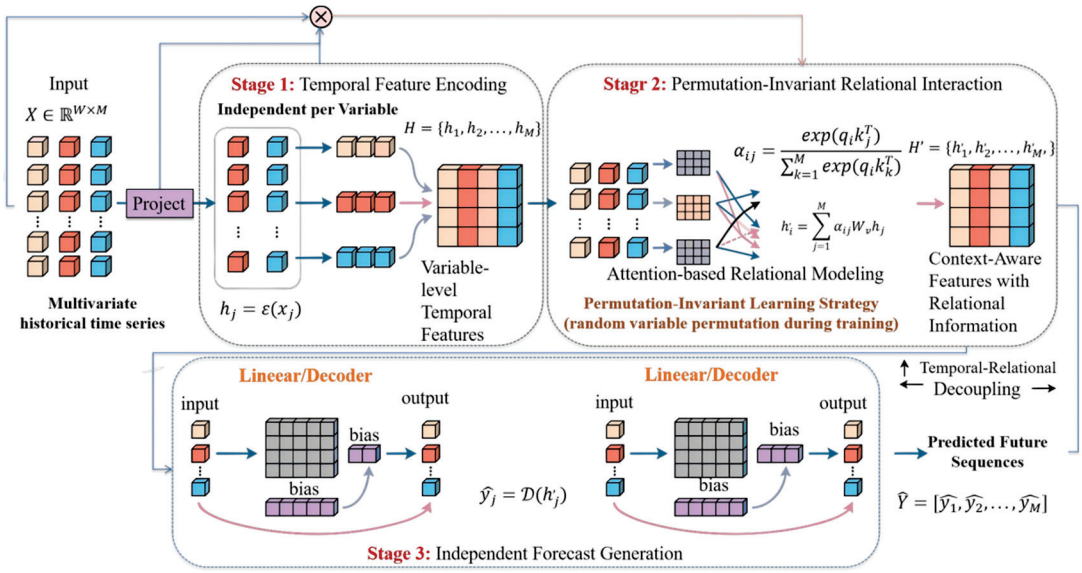


Figure 3. Structure diagram of the CPiRi method.

3.3. Core Model Architecture

CPiRi adopts a temporal–relational decoupled architecture for modeling. The model first extracts temporal feature representations for each variable, then captures inter-variable interaction information through a relational modeling module, and finally generates future sequence predictions. The overall architecture consists of three main modules: a temporal feature encoding module, a relational interaction modeling module, and a prediction generation module.

3.3.1. Temporal Feature Encoding

Given the historical observation sequence $X \in \mathbb{R}^{W \times M}$, where W denotes the length of the historical time window and M represents the number of variables, the time series corresponding to the j-th variable can be expressed as:

$$x_j = [x_{1,j}, x_{2,j}, \ldots, x_{W,j}] \tag{5}$$

The temporal encoding module independently models each variable sequence to extract its temporal dynamic features:

$$h_j = \mathcal{E}(x_j) \tag{6}$$

where $\mathcal{E}(\cdot)$ denotes the temporal encoding function, $h_j \in \mathbb{R}^d$ represents the temporal feature of variable j, and d is the feature dimension.

After encoding all variables, the variable-level temporal feature set can be obtained as:

$$H = \{h_1, h_2, \ldots, h_M\} \tag{7}$$

This feature set characterizes the dynamic evolution patterns of each variable within the historical time window.

https://doi.org/10.3390/fire9070267

3.3.2. Relational Interaction Modeling

After obtaining the temporal feature representations, CPiRi captures inter-variable interactions through a relational modeling module.

Let the set of variable features be:

$$H = \{h_1, \ldots, h_M\} \tag{8}$$

The relational module learns dependencies among variables via an attention mechanism:

$$h_i' = \sum_{j=1}^{M} \alpha_{ij} W_v h_j \tag{9}$$

where the attention weights are defined as:

$$\alpha_{ij} = \frac{\exp\left(q_i k_j^T\right)}{\sum_{k=1}^{M} \exp\left(q_i k_k^T\right)} \tag{10}$$

and

$$q_i = W_q h_i \tag{11}$$

$$k_j = W_k h_j \tag{12}$$

Here, W_q, W_k, W_v are learnable parameter matrices, and α_{ij} represents the attention weight indicating the degree of relevance from variable i to variable j.

Through the above relational modeling process, each variable feature can integrate information from other variables, thereby forming a new contextual representation:

$$H' = \{h_1', h_2', \ldots, h_M'\} \tag{13}$$

This representation simultaneously encodes both temporal features and inter-variable relational information.

3.3.3. Prediction Generation

After obtaining the updated variable representations, the model generates future sequences through a prediction module.

For the j-th variable, the prediction result is given by:

$$\hat{y}_j = \mathcal{D}(h_j\prime) \tag{14}$$

where $\mathcal{D}(\cdot)$ denotes the prediction decoder, and $\hat{y}_j \in \mathbb{R}^H$.

The final prediction can be expressed as:

$$\hat{Y} = [\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_M] \tag{15}$$

which corresponds to the predicted values of all variables over the next H time steps.

3.4. Permutation-Invariant Training Strategy

In multivariate time-series forecasting tasks, the ordering of variables is often arbitrary. For instance, in sensor networks, financial indicators, or industrial monitoring systems, variable indices may change due to device replacement, system upgrades, or data reorgani-

https://doi.org/10.3390/fire9070267

zation. If a model relies on a fixed variable ordering during training, its performance is likely to degrade significantly when the variable order changes.

To address this issue, CPiRi introduces a permutation-invariant training strategy during the training phase. This strategy exposes the model to different variable permutations through random shuffling, enabling it to learn relational representations that are independent of variable ordering.

Let the number of variables be M, and denote the set of all permutations of variable indices as Π_M. At each training iteration, a permutation $\sigma \sim \Pi_M$ is randomly sampled. The input sequence and the target sequence are then simultaneously permuted as X^σ, Y^σ, respectively, where X^σ denotes the permuted input sequence and Y^σ denotes the corresponding permuted target sequence.

The model prediction is given by:

$$\hat{Y}^\sigma = \mathrm{CPIRI}(X^\sigma) \tag{16}$$

The training objective is defined by minimizing the prediction error:

$$\mathcal{L} = \mathcal{L}\left(\hat{Y}^\sigma, Y^\sigma\right) \tag{17}$$

Common loss functions include Mean Squared Error (MSE) and Mean Absolute Error (MAE). Furthermore, the overall training objective can be formulated as:

$$\min_{\Theta} \mathbb{E}_{(X,Y)\sim D, \sigma\sim\Pi_M}\left[\mathcal{L}\left(\hat{Y}^\sigma, Y^\sigma\right)\right] \tag{18}$$

where θ denotes the model parameters and $\mathcal{D}$ represents the training dataset.

Through this training strategy, the model is encouraged to learn relational structures based on variable content rather than relying on fixed variable indices, thereby improving its generalization ability under variable reordering. The overall training pipeline of CPiRi is illustrated in Figure 4.

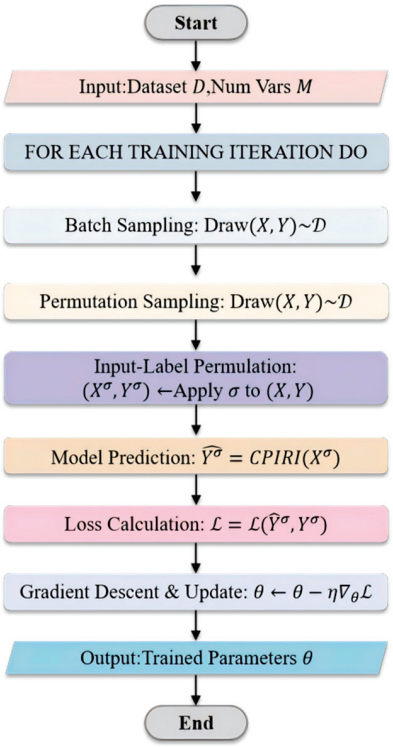


Figure 4. CPIRI Training Process.

https://doi.org/10.3390/fire9070267

3.5. Permutation-Invariant Property Analysis

To further demonstrate the robustness of the CPiRi framework under variable order variations, we provide a theoretical analysis of its permutation-invariant properties.

Let the multivariate feature representation be defined as:

$$H = \{h_1, h_2, \ldots, h_M\} \tag{19}$$

where h_i denotes the feature vector of the i-th variable. For any permutation of variable indices $\sigma \in \Pi_M$, the permutation operation can be defined as:

$$\sigma(H) = \left\{h_{\sigma(1)}, h_{\sigma(2)}, \ldots, h_{\sigma(M)}\right\} \tag{20}$$

If a function $f(\cdot)$ satisfies:

$$f(\sigma(H)) = \sigma(f(H)) \tag{21}$$

then the function is said to be permutation equivariant. This property implies that when the input variable order changes, the output undergoes the same permutation, while the underlying relational structure among variables remains unchanged.

In the CPiRi framework, the relational interaction module aggregates information across variables via an attention mechanism. According to the formulation in Section 3.3, the updated variable representation is given by:

$$h_i' = \sum_{j=1}^{M} \alpha_{ij} W_v h_j \tag{22}$$

where the attention weights are defined as:

$$\alpha_{ij} = \frac{\exp\left(q_i k_j^T\right)}{\sum_{k=1}^{M} \exp\left(q_i k_k^T\right)} \tag{23}$$

Since the attention computation depends solely on the similarity between variable features rather than their absolute indices, a permutation of the input variables results in a corresponding permutation of the attention weight matrix, thereby preserving the overall relational structure.

Furthermore, during training, a random variable permutation strategy is introduced, allowing the model to learn under diverse variable arrangements. In this way, the model is encouraged to capture relational patterns based on variable content rather than relying on fixed variable ordering. Consequently, after training, CPiRi can maintain stable predictive performance under scenarios such as variable reordering, variable addition, or missing variables.

In summary, by combining the permutation-equivariant structure of the relational module with the random permutation strategy during training, CPiRi is able to learn order-invariant relational representations, thereby enhancing both the generalization ability and robustness in multivariate time-series forecasting tasks.

4. Case Study

4.1. Data Source

To thoroughly uncover the multi-scale coupled characteristics embedded in gas concentration time-series data and to validate the channel permutation invariance and generalization capability of the proposed CPiRi model under dynamic mining environments, this study selects monitoring data covering diverse gas occurrence conditions and min-

https://doi.org/10.3390/fire9070267

ing intensities as the foundation for model development. Leveraging long-term research collaborations between our team and multiple coal enterprises in China, and considering variations in geological complexity and non-stationarity of gas emission, 15 mining faces with strong channel heterogeneity were selected from five representative coal mines located in major coal-producing regions, including Henan, Anhui, Guizhou, and Heilongjiang, as data collection sites. The basic information of each coal mine can be found in the Supplementary Materials. To avoid data distortion caused by sensor drift or adjustments in mining operations and to ensure that the data faithfully reflect the dynamic evolution of gas concentration, all monitoring periods were uniformly selected from the third quarter of 2025, during which mining activities were stable and ventilation systems operated under normal conditions.

The core experimental dataset in this study is derived from a fully mechanized mining face in a coal and gas outburst-prone mine located in Pingdingshan, Henan Province. This site is equipped with an intelligent gas monitoring and early warning system jointly developed by our team. The system synchronously collects multi-channel time-series data—including gas concentration, wind speed, temperature, and pressure—at a sampling interval of 30 s. To construct a dataset suitable for multi-step forecasting tasks, a sliding window technique is applied to segment the raw time series into samples. Following strict temporal ordering, the generated samples are split into training, validation, and test sets in a ratio of 7:2:1, yielding approximately 120,000 valid samples. The statistical characteristics and physical meanings of all monitored variables are summarized in Table 1.

Table 1. Data attributes of each measurement point.

Point Location	Station Description	Index	Maximum Value	Minimum Value
Mix_Gas (label)	Mixed methane concentration in air inlet	$\%CH_4$	0.76	0
En_Gas	Methane concentration in inlet and return air	$\%CH_4$	0.62	0
Gas_1	Methane concentration on the wind side of the tunnel under drilling	$\%CH_4$	0.82	0.09
Gas_2	Methane concentration in inlet face	$\%CH_4$	0.52	0
CO_1	Carbon monoxide concentration on the wind side of drilling in the air inlet	ppm	5	0
CO_2	Carbon monoxide concentration of belt head in air inlet	ppm	6	0
Wind Speed	Wind speed of inlet and return air	m/s	1.4	0.2
Dust	Driving face dust in air inlet lane	mg/m^3	0	0
En_T	Inlet and return air temperature	°C	24.9	13.5
Gas Extraction	Mixed instantaneous flow rate in air inlet duct condition	m^3	13.15	0
Smoke	Smoke on the downwind side of the belt head in air inlet lane	mg/m^3	0	0

During data acquisition at the front end of the intelligent sensing system, built-in robust filtering techniques have already been applied to effectively suppress abnormal outliers caused by transient sensor interruptions or electromagnetic interference. Statistical analysis shows that the completeness of the raw dataset reaches 99.8%. For the small number of data points flagged as uncertain by the system, a moving-average method based on adjacent timestamps is employed for smoothing correction, ensuring temporal continuity and reliability while preserving the intrinsic dynamic evolution patterns as much as possible. Considering the significant differences in numerical scales among input features, min–max normalization is applied to independently rescale each feature to the range [0, 1], eliminating the influence of physical units and aligning with the characteristics

https://doi.org/10.3390/fire9070267

of deep learning activation functions. This normalization step serves as a crucial foundation for enabling the model to treat different physical variables equitably and effectively capture their intrinsic coupling relationships.

4.2. Ablation Study

To systematically evaluate the effectiveness of the core design components of the CPiRi framework, a series of ablation experiments is conducted. Specifically, to assess the contributions of architectural decoupling, training regularization, and spatial module design, ablation variants are constructed along these three dimensions. By selectively retaining or replacing specific components within a single dimension, a total of seven ablation variants are designed and compared with the full model, allowing quantitative assessment of each component's independent contribution to prediction accuracy and permutation robustness.

All experiments adopt the same input length L = 64 and prediction horizon T = 12. The evaluation metrics include Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). Model hyperparameters are uniformly set as follows: hidden dimension of 128, 4 attention heads, dropout rate of 0.1, Adam optimizer with a learning rate of 0.001, and 50 training epochs, with early stopping applied to prevent overfitting.

To ensure statistical reliability and explicitly evaluate the stability of the proposed framework, both the full model and all ablation variants were independently executed five times under identical settings. The evaluation metrics include Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). All results are reported in the format of "mean $\pm$ standard deviation" to reflect experimental variance. Furthermore, to verify whether the performance improvements of CPiRi are statistically significant rather than coincidental due to random initialization, we conducted paired *t*-tests between the full CPiRi model and each variant across the five independent runs. Detailed configurations are provided in Table 2, and the comprehensive statistical prediction results under different experimental settings are summarized in Table 3.

Table 2. Configuration of ablation experiments.

Model Variant	Encoder Status	Training Strategy	Space Module Design
CPiRi (Full)	Freeze	Channel Shuffle	Self-attention
Variant A: Micro-coupling decoupling	Fine Tuning	Channel Shuffle	Self-attention
Variant B: Fixed training	Freeze	Permanent Order	Self-attention
Variant C: No spatial interaction	Freeze	Channel Shuffle	/
Variant D: Random encoder	Random Initialization	Channel Shuffle	Self-attention
Variant F: Average pooling	Freeze	Channel Shuffle	Average pooling
Variant G: Fully connected layer	Freeze	Channel Shuffle	fully connected network

The statistical ablation results provide profound insights into the functional roles of different components within the CPiRi framework, while explicitly validating the model's stability. As shown in Table 3, the full CPiRi model maintains the lowest standard deviations across all tasks, underscoring its robustness. The paired *t*-tests confirm that the performance advantages of CPiRi over all variants are statistically significant, with all exact *p*-values being well below the 0.05 threshold, as detailed in Table 3. The comparison between Variant A and the full model indicates that freezing the pretrained encoder prevents overfitting to specific mine-level data distributions during fine-tuning, thereby preserving the general representation capability learned from large-scale time-series data. More crucially, the performance gap observed in Variant B between standard testing and permutation testing highlights the necessity of the channel-shuffling strategy. Without this

https://doi.org/10.3390/fire9070267

regularization, Variant B not only suffers a severe drop in average accuracy during permutation tests but also exhibits a massive surge in standard error. This demonstrates that the model tends to memorize fixed channel orders, leading to severe performance degradation and extreme instability when input order changes—sometimes even performing worse than Variant C, which completely removes inter-channel interactions. This suggests that under improper training strategies, complex spatial modules may instead become a source of model fragility. The performance degradation of Variant C confirms the indispensable role of inter-channel relationship modeling in gas concentration prediction. The coupling of gas emission among monitoring points such as mining faces and return airways is governed by physical mechanisms, which cannot be captured through independent modeling. Furthermore, the statistical inferiority of Variants F and G compared to the full model further demonstrates the advantages of the self-attention mechanism: average pooling is limited by its inability to model nonlinear coupling relationships, while fully connected networks lack permutation equivariance, as evidenced by its significantly widened standard deviation under channel reordering. These observations indirectly validate the superiority of self-attention in both aggregation flexibility and permutation-equivariant properties. The results of Variant D show that randomly initialized encoders fail to provide sufficient temporal priors. Even when combined with spatial modules and shuffling strategies, prediction accuracy still deteriorates significantly, underscoring the fundamental importance of pretrained weights. Overall, the performance gains of CPiRi stem from the coordinated design of its components: the frozen encoder ensures stable temporal feature extraction, the channel-shuffling strategy enforces content-based relational learning rather than position-based dependency, and the self-attention mechanism provides the permutation-equivariant structure necessary to achieve this objective.

Table 3. Prediction results for each experimental configuration.

Model Variant	MAE	RMSE	*p*-Value	Permutation Test MAE	Permutation Test RMSE	*p*-Value
CPiRi (Full)	6.12 ± 0.08	8.03 ± 0.11	-	6.15 ± 0.09	8.07 ± 0.25	-
Variant A: Micro-coupling decoupling	6.71 ± 0.14	8.82 ± 0.18	0.003	7.19 ± 0.21	9.44 ± 0.25	0.002
Variant B: Fixed training	6.44 ± 0.12	8.46 ± 0.15	0.014	9.94 ± 0.85	13.05 ± 1.12	<0.001
Variant C: No Spatial interaction	7.98 ± 0.15	10.49 ± 0.22	<0.001	8.11 ± 0.18	10.63 ± 0.24	<0.001
Variant D: Random encoder	9.67 ± 0.28	12.71 ± 0.35	<0.001	10.02 ± 0.32	13.14 ± 0.41	<0.001
Variant F: Average pooling	7.21 ± 0.16	9.48 ± 0.20	0.002	7.45 ± 0.19	9.79 ± 0.22	0.002
Variant G: Fully connected layer	6.94 ± 0.13	9.13 ± 0.17	0.005	8.61 ± 0.45	11.31 ± 0.58	<0.001

Note: The exact *p*-values reported represent the more conservative (higher) value obtained from paired *t*-tests on both MAE and RMSE metrics, comparing each variant against the full CPiRi model across the five independent runs. All values confirm statistical significance ($p < 0.05$).

4.3. Analysis of Prediction Results

To comprehensively evaluate the effectiveness of the CPiRi model in gas concentration prediction, systematic experiments were conducted across 15 real-world excavation faces covering several typical coal-producing regions in China. First, single-step prediction visu-

https://doi.org/10.3390/fire9070267

alizations were utilized to intuitively demonstrate the performance gap between CPiRi and four mainstream baseline models—PatchTST, DLinear, iTransformer, and Crossformer—in fitting the dynamic evolution of gas concentrations. Subsequently, quantitative comparisons were performed against these baselines in multi-step prediction tasks to verify the performance advantages of CPiRi across various prediction scenarios. Figure 5 illustrates the comparison curves between the single-step prediction results of CPiRi and the baseline models against the ground truth across the 15 test faces.

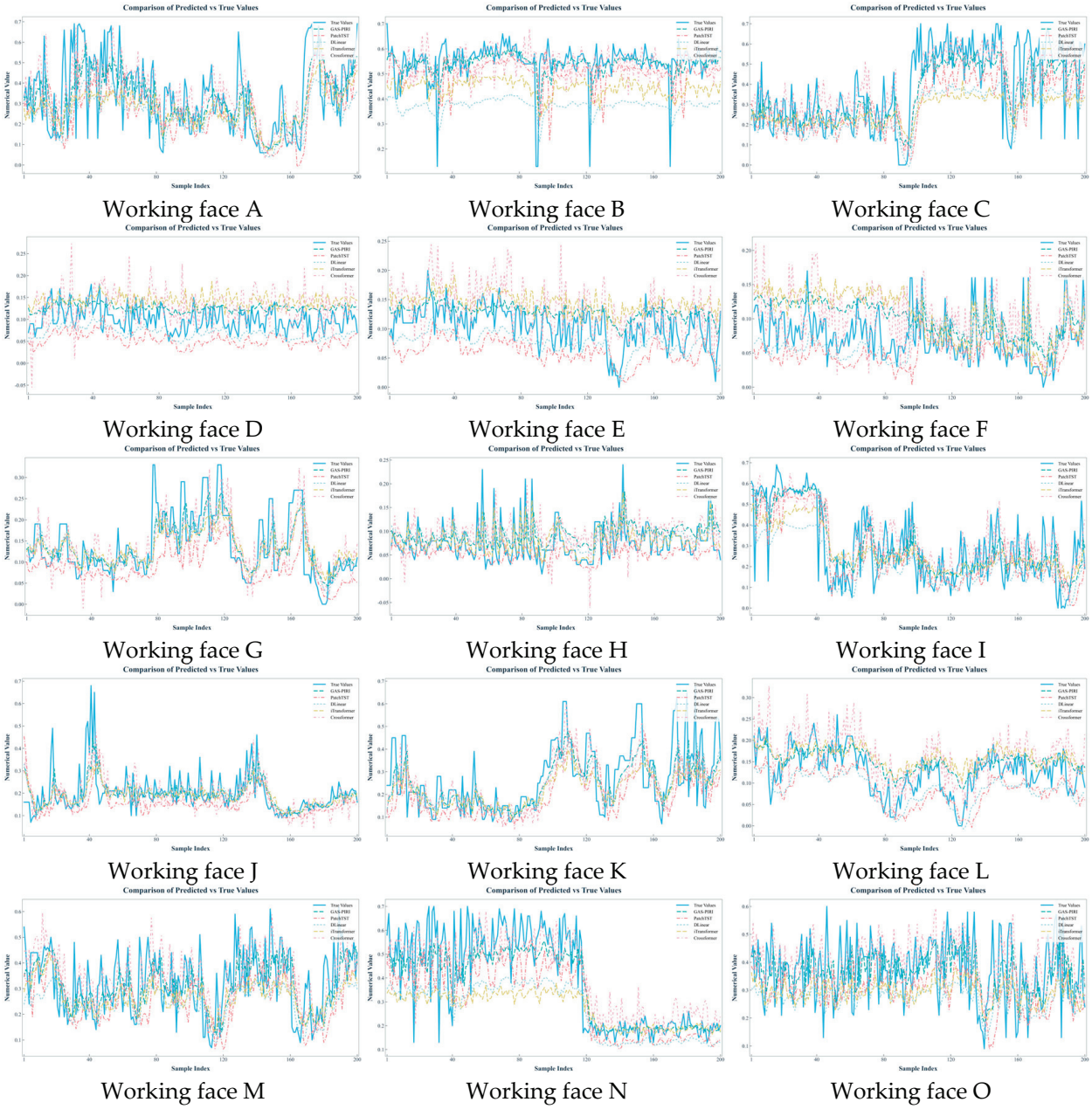


Figure 5. Prediction results of each model.

In excavation faces where gas concentration fluctuates drastically, the performance disparities among models are particularly pronounced. Taking Working Face B as an example, the area experienced four consecutive rapid declines; CPiRi successfully limited the prediction errors for the two peak magnitudes to within 5%. While iTransformer and Crossformer managed to capture the primary oscillatory trends, their predicted values near the peaks were generally lower than the actual values. Conversely, PatchTST and DLinear,

https://doi.org/10.3390/fire9070267

due to their lack of inter-channel coupling modeling, exhibited a significant lag in response within mutation regions, with peak prediction delays of approximately 2 to 3 sampling points and an average relative error exceeding 10%. This discrepancy indicates that when the historical information of a target channel is insufficient to support an accurate prediction, channel-independent models fail to leverage the early responses of other correlated channels for calibration. In working faces with relatively stable concentration changes, the performance differences among models are reflected in the fitting precision of subtle local fluctuations. For instance, at Working Face D, CPiRi maintained an Mean Absolute Error of below 4.5% within the stable interval. While the prediction curves of iTransformer and Crossformer followed the overall correct trend, they exhibited over-smoothing at local fluctuations, showing insufficient sensitivity to small-scale responses. PatchTST and DLinear displayed trend-level deviations from the ground truth during certain periods, suggesting that even in stable intervals, cross-channel information plays a proactive role in maintaining prediction accuracy. Synthesizing the single-step prediction results from all 15 working faces, CPiRi demonstrated a highly effective fitting effect across all test scenarios, showing the highest degree of overlap with the ground truth, the most accurate capture of mutation features, and robust performance against noise interference.

To verify the advantages of CPiRi in multi-step prediction tasks, five representative mining faces were selected to compare CPiRi with both channel-independent and channel-dependent models. All models utilized identical input lengths (L = 64) and prediction lengths (T = 12), with MAE and RMSE selected as evaluation metrics. The results are summarized in Table 4.

Table 4. Multi-step Prediction Results of Each Model.

Model	CPiRi		PatchTST		DLinear		iTransformer		Crossformer	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Test working face I	6.52	8.56	7.25	9.54	7.99	10.51	6.81	8.95	7.11	9.34
Test working face II	6.87	9.02	7.64	10.06	8.41	11.08	7.18	9.44	7.49	9.86
Test working face III	6.68	8.79	7.44	9.80	8.20	10.80	6.99	9.20	7.30	9.61
Test working face IV	7.07	9.30	7.86	10.36	8.67	11.41	7.39	9.73	7.72	10.16
Test working face V	7.32	9.64	8.15	10.74	8.99	11.84	7.66	10.09	8.01	10.55

As shown in Table 4, CPiRi achieved the optimal MAE and RMSE across all five test faces. Compared to iTransformer, CPiRi's relative improvement in MAE stabilized between 4.3% and 4.4%, with a similar magnitude of improvement in RMSE. This advantage remained highly consistent across working faces with varying geological conditions and excavation intensities, verifying the model's robust generalization capability. Compared to channel-independent models, CPiRi's superiority was even more significant. The relative improvement in MAE was approximately 10.0% to 10.2% compared to PatchTST, and reached 18.4% to 18.6% compared to DLinear. Further analysis of the baseline performances reveals that while channel-independent models possess inherent robustness to channel ordering in gas concentration tasks, they significantly underperform compared to models capable of inter-channel interaction modeling, as they neglect the gas emission coupling relationship between monitoring points such as the excavation face and the return air lane. For example, the average MAE for PatchTST across the five faces was 7.67, which is approximately 11.2% higher than that of CPiRi; the average MAE for DLinear reached 8.45, roughly

https://doi.org/10.3390/fire9070267

22.6% higher than CPiRi. These results indicate that in multivariate gas concentration prediction, the information embedded in inter-channel relationships is indispensable, and relying solely on independent channel modeling fails to reach peak performance. Channel-dependent models attained superior performance over channel-independent counterparts by explicitly modeling channel relationships. iTransformer's average MAE across the five faces was 7.21, a 6.0% reduction compared to PatchTST, confirming the value of channel relationship modeling. Crossformer also showed a notable improvement over PatchTST with an average MAE of 7.53. However, such models rely on an interaction mechanism with fixed channel indices, and their performance is predicated on the assumption that the channel order in the training data remains invariant. In practical mining environments, sensors frequently undergo renumbering or position adjustments due to maintenance, malfunctions, or system upgrades; the performance of these models in dynamic sensor networks will be further verified in subsequent permutation tests.

Through the synergistic design of a spatio-temporal decoupling architecture and a channel-shuffling strategy, CPiRi retains the capability for channel relationship modeling while eliminating dependence on a fixed sequence. Compared to iTransformer, CPiRi reduced the average MAE across the five faces by approximately 4.4%. This advantage primarily stems from two architectural distinctions: first, the frozen pre-trained encoder provides a more stable and robust temporal feature extraction capability, preventing the encoder from over-fitting to the specific distribution of a single mine during fine-tuning; second, the channel-shuffling strategy forces the spatial module to learn universal relational reasoning based on data content rather than memorizing fixed sequence indices. This more fundamental understanding of relationships yields performance gains even in routine prediction tasks.

5. Discussion

This study addresses the pragmatic challenge of channel order uncertainty in coal mine gas concentration prediction, arising from the dynamic nature of sensor networks. We propose CPiRi, a prediction network based on Channel Permutation-Invariant Relation Interaction. Through the synergistic design of a spatio-temporal decoupling architecture and a channel-shuffling regularization strategy, the model systematically reconciles the inherent contradiction between channel-dependent and channel-independent methodologies. It preserves the robustness of channel-independent models against sequential variations while inheriting the capacity of channel-dependent models to capture cross-monitoring point gas emission coupling relationships.

Experimental results demonstrate that CPiRi achieves superior prediction accuracy across multiple real-world mining datasets. Ablation studies on core modules confirm that the frozen pre-trained encoder, the channel-shuffling strategy, and the self-attention spatial interaction are indispensable and exhibit significant synergistic effects. Permutation tests further reveal that an equivariant architecture alone is insufficient to achieve true channel permutation invariance; it must be coupled with channel-shuffling training to compel the model to bypass "positional memory shortcuts" in favor of learning content-driven semantic relationships. Collectively, these findings establish the core contribution of CPiRi: transforming the theoretical concept of channel permutation invariance into a realizable framework, thereby providing a reliable technical pathway for gas concentration prediction in dynamic mining environments with frequent sensor reconfigurations [23].

A comparative analysis with existing gas concentration prediction methods reveals two primary categories of traditional approaches. The first includes analytical methods based on physical models, which, despite their interpretability, struggle to adapt to the dynamic changes in excavation environments. The second comprises data-driven deep

https://doi.org/10.3390/fire9070267

learning models. Within this category, channel-dependent models can capture inter-channel relationships but suffer performance collapse during sensor renumbering or modification due to over-fitting to fixed channel orders [24]. Conversely, channel-independent models suffer from limited precision due to their neglect of cross-channel coupling. The advancement of CPiRi does not stem merely from an increase in model capacity, but from a targeted design addressing the core issue of channel order uncertainty. Traditional channel-dependent models treat channel order as a fixed prior, essentially assuming a static binding between channel indices and physical sensors—a fundamental mismatch with the dynamic realities of sensor replacement and system upgrades in mines. CPiRi's adaptive relational modeling effectively bridges this gap. This progress represents a deeper understanding of the dynamic characteristics of gas-monitoring networks: while traditional methods attempt to fit varying configurations with fixed-order models, CPiRi enables the model to evolve from "memorizing sequences" to "understanding relations," aligning more closely with the underlying physical nature of spatial gas emission correlations [25].

Furthermore, it is imperative to establish a rigorous boundary regarding the model's current applicability. While CPiRi successfully resolves the issue of channel order uncertainty through permutation-invariant modeling, channel permutation is not strictly equivalent to all dynamic operation and maintenance anomalies. Currently, CPiRi maintains a fixed input dimension M, effectively addressing scenarios involving sensor re-indexing, replacement, and channel shuffling. However, practical mining environments also face dynamic scaling of the monitoring network (sensor additions and removals) which alters the dimensional structure, as well as physical data corruption (missing values and sensor drift). The current architecture does not dynamically adapt to varying input sizes, nor has its robustness against sensor drift been fully verified. In actual mines, anomalies such as sensor drift or transmission interruptions are frequent [26].

Building on this foundation, future research may expand from single-point prediction to joint spatio-temporal forecasting, combining permutation-invariant concepts with Graph Neural Networks (GNNs) to capture the spatio-temporal correlations between multiple measuring points. This would enable collaborative prediction across critical locations such as excavation faces and return air lanes. Simultaneously, the focus could evolve from concentration prediction toward risk situational awareness. Through multi-task learning, the system could simultaneously output derivative indicators such as gas emission rates and over-limit probabilities, providing a richer decision-making basis for differentiated early-warning strategies. Long-term, embedding CPiRi into intelligent coal mine sensing systems to construct a "perception–prediction–warning–control" closed loop promises to shift gas management from reactive response to proactive prevention, providing more robust technical support for mine production safety [27].

6. Conclusions

This study proposes CPiRi, a gas concentration prediction network based on channel permutation-invariant relation interaction, achieving precise and robust predictions within dynamic sensor network environments. The innovation of this method lies in identifying channel order uncertainty—primarily caused by sensor renumbering, hardware replacement, and channel routing adjustments—as a fundamental barrier limiting the applicability of existing methods [28,29]. To address this, we constructed a spatio-temporal decoupling architecture to independently extract temporal features from each monitoring point while introducing a permutation-equivariant spatial module to model gas emission relationships based on data content. This research introduces a channel-shuffling regularization training strategy, breaking the limitations of traditional models that rely on fixed-order indices and achieving near-invariant prediction performance under random channel permuta-

https://doi.org/10.3390/fire9070267

tions. Evaluated on real-world datasets covering several typical Chinese coal mines, the method demonstrated superior prediction accuracy and permutation robustness. This not only validates the design but also provides an innovative technical pathway for gas monitoring scenarios where sensor configurations frequently change. Through systematic experimentation, this paper concludes:

Channel order uncertainty is the core bottleneck for the performance degradation of existing gas concentration prediction methods in dynamic networks. Underground sensors frequently face renumbering or replacement due to maintenance and upgrades. Channel-dependent models suffer sharp performance declines when the input order deviates from the training stage, while channel-independent models sacrifice accuracy by ignoring cross-channel relationships. This trade-off between relationship modeling and sequential adaptability is a fundamental contradiction that existing methods fail to reconcile.

CPiRi systematically resolves this contradiction through the synergy of spatio-temporal decoupling and channel-shuffling regularization. The architecture separates a frozen pre-trained univariate encoder from a permutation-equivariant spatial module: the former extracts robust temporal features independently (inheriting the stability of channel-independent models), while the latter utilizes self-attention to mine semantic relationships based on content rather than indices. Channel-shuffling during training forces the spatial module to learn universal relational reasoning, allowing the model to accurately identify physical correlations between monitoring points regardless of their order. Multidimensional experiments verify the effectiveness, robustness, and utility of CPiRi. The complete model achieved optimal MAE and RMSE across multiple real-world mine datasets. Ablation studies confirmed that the pre-trained encoder, shuffling strategy, and self-attention module are mutually reinforcing. This research adapts and deepens the application of these concepts within the gas concentration prediction scenario [30]. While CPiRi represents a paradigm shift from "memorizing order" to "understanding relations," future work must continue to refine domain-adaptive fine-tuning, dynamic fusion mechanisms, and the injection of physical knowledge to overcome current dependencies and further enhance the "perception-to-control" technical chain in coal mine safety.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/fire9070267/s1. Table S1: Key Parameters of the Prediction Model for Coal Mine A. Table S2: Key Parameters of the Prediction Model for Coal Mine B. Table S3: Key Parameters of the Prediction Model for Coal Mine C. Table S4: Key Parameters of the Prediction Model for Coal Mine D. Table S5: Key Parameters of the Prediction Model for Coal Mine E. Table S6: Key Parameters of the Prediction Model for Coal Mine F. Table S7: Key Parameters of the Prediction Model for Coal Mine G. Table S8: Key Parameters of the Prediction Model for Coal Mine H. Table S9: Key Parameters of the Prediction Model for Coal Mine I. Table S10: Key Parameters of the Prediction Model for Coal Mine J. Algorithm S1: Training Pipeline of the CPiRi Framework.

Author Contributions: Conceptualization, Y.Y. and Y.Z.; methodology, Y.Y.; software, Y.Z.; validation, Y.Z.; formal analysis, Y.Z.; investigation, J.H.; resources, Y.Y.; data curation, Y.Y.; writing—original draft preparation, Y.Y.; writing—review and editing, J.H.; visualization, Y.Y.; supervision, Y.Z.; project administration, Y.Y.; funding acquisition, J.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are available on request from the corresponding author due to confidentiality agreements signed with our collaborative partners, which prohibit the public sharing of the raw dataset.

Conflicts of Interest: The authors declare no conflicts of interest.

https://doi.org/10.3390/fire9070267

References

1. Li, S.; Wang, X.; Zhang, Y.; Lu, C.; Xu, N. Research on multi-state risk prediction and evidence diagnosis of coal mine gas explosion integrating fuzzy dynamic logic gates and DBN. *J. Loss Prev. Process Ind.* **2026**, *103*, 0950–4230. [CrossRef]
2. Li, S.; Xu, N.; Lu, C.; Zhang, Y.; Xu, K. Coal mine methane risk prediction method based on data characteristics and decomposition-integration strategy. *Comput. Ind. Eng.* **2026**, *211*, 111631. [CrossRef]
3. Xing, W.; Yang, Y.; Zhang, Y.; Yang, Y. Gas concentration prediction based on SSA algorithm with CNN-BiLSTM-attention. *Sci. Rep.* **2025**, *15*, 34214. [CrossRef] [PubMed]
4. Zhou, J. Research on Disaster Early Warning Method Based on Gas Emission Law of Working Face. Master's Thesis, Xi'an University of Science and Technology, Xi'an, China, 2019.
5. Nie, Y.; Nguyen, N.H.; Sinthong, P.; Kalagnanam, J. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. *arXiv* **2023**, arXiv:2211.14730. [CrossRef]
6. Zhang, T.; Song, S.; Li, S.; Ma, L.; Pan, S.; Han, L. Research on Gas Concentration Prediction Models Based on LSTM Multidimensional Time Series. *Energies* **2019**, *12*, 161. [CrossRef]
7. Wu, R.M.X.; Shafiabady, N.; Zhang, H.; Lu, H.; Gide, E.; Liu, J.; Charbonnier, C.F.B. Comparative study of ten machine learning algorithms for short-term forecasting in gas warning systems. *Sci. Rep.* **2024**, *14*, 21969. [CrossRef] [PubMed]
8. Huang, Y.; Fan, J.; Yan, Z.; Li, S.; Wang, Y. A Gas Concentration Prediction Method Driven by a Spark Streaming Framework. *Energies* **2022**, *15*, 5335. [CrossRef]
9. Fan, J.; Huang, Y.; Yan, Z.; Li, C.; Wang, C.; He, Y. Gas concentration prediction based on ARIMA-SVM hybrid model. *Ind. Mine Autom.* **2022**, *48*, 134–139. [CrossRef]
10. Li, C.; Fang, X.; Yan, Z.; Huang, Y.; Liang, M. Research on Gas Concentration Prediction Based on the ARIMA-LSTM Combination Model. *Processes* **2023**, *11*, 174. [CrossRef]
11. Liu, C.; Zhang, A.; Xue, J.; Lei, C.; Zeng, X. LSTM-Pearson Gas Concentration Prediction Model Feature Selection and Its Applications. *Energies* **2023**, *16*, 2318. [CrossRef]
12. Wang, Y.; Zhang, L.; Yan, Z.; Deng, J.; Huang, Y.; Qin, Z.; Cao, Y.; Wang, Y. TCN–Transformer Spatio-Temporal Feature Decoupling and Dynamic Kernel Density Estimation for Gas Concentration Fluctuation Warning. *Fire* **2025**, *8*, 175. [CrossRef]
13. Su, P.; Zhang, J.; Guo, L. A PINN-LSTM Model for Real-Time Gas Concentration Prediction in Mine Conveyor Belt Fires. *Fire* **2025**, *8*, 450. [CrossRef]
14. Tan, B.; Yang, K.; Sui, L.; Zuo, Y.; Gao, S.; Tang, S.; Jia, J. Research on gas concentration prediction model based on xLSTM-Informer. *Ind. Mine Autom.* **2025**, *51*, 81–89. [CrossRef]
15. Hao, J.; Chen, Y.; Lv, C.; Yuan, X.; Wang, H.; Zhang, Q.; Dong, Y. A visual analytics framework for unveiling multi-scale patch decomposition in transformer-based time-series forecasting. *J. Vis.* **2026**, *29*, 137–152. [CrossRef]
16. Lee, H.; Kim, Y.; Jung, S. An Effective Summary Preprocessing Method for Time Series Forecasting with Multiple Temporal Granularities. *IEEE Access* **2025**, *13*, 51277–51286. [CrossRef]
17. Ding, K.; Fan, F.; Hou, C.; Wang, Z.; Wang, L.; Yang, Z.; Zhan, J. TimeMosaic: Temporal Heterogeneity Guided Time Series Forecasting via Adaptive Granularity Patch and Segment-wise Decoding. *Proc. AAAI Conf. Artif. Intell.* **2026**, *40*, 20790–20798. [CrossRef]
18. Xu, N.; Li, S.; Liu, J.; Lu, C.; Wang, X. Research on methane concentration prediction method based on non-steady-state feature coupling interaction and structural perception loss. *Reliab. Eng. Syst. Saf.* **2026**, *272*, 112530. [CrossRef]
19. Li, S.; Xu, N. A spatio-temporal graph attention prediction method for coal spontaneous combustion temperature integrating time-frequency domain lag features. *Eng. Appl. Artif. Intell.* **2026**, *167*, 113773. [CrossRef]
20. Xu, N.; Li, S.; Xu, K.; Lu, C. Research on methane Hazard interval prediction method based on hybrid "model-data"driven strategy. *Appl. Energy* **2025**, *377*, 124613. [CrossRef]
21. Yan, X.; Zhao, X.; Chi, H. FreqMamba: Frequency-aware multi-graph fusion with Mamba for traffic flow prediction. *Knowl.-Based Syst.* **2026**, *347*, 116316. [CrossRef]
22. Lu, C.; Li, S.; Xu, N.; Zhang, Y.; Qin, Y. Hybrid-driven risk assessment methodology for coal and gas outburst: Integration of complex network, disaster mechanism, and multi-level fusion modeling. *Process Saf. Environ. Prot.* **2025**, *199*, 107226. [CrossRef]
23. Huang, J.; Liu, L.; Zhao, H.; Li, T.; Li, B. RUL-Mamba: Mamba-based remaining useful life prediction for lithium-ion batteries. *J. Energy Storage* **2025**, *120*, 116376. [CrossRef]
24. Peng, F.Z.; Wang, C.; Chen, T.; Schussheim, B.; Vincoff, S.; Chatterjee, P. PTM-Mamba: A PTM-aware protein language model with bidirectional gated Mamba blocks. *Nat. Methods* **2025**, *22*, 945–949. [CrossRef] [PubMed]
25. Cai, M.; Zhang, T.; Liu, P.; Zhou, Q. APMoE-Net: Fourier amplitude-phase joint enhancement and MoE compensation for low-light image enhancement. *Expert Syst. Appl.* **2026**, *298*, 129664. [CrossRef]
26. Liu, C.; Wang, E.; Li, Z.; Zang, Z.; Li, B.; Yin, S.; Zhang, C.; Liu, Y.; Wang, J. Research on multi-factor adaptive integrated early warning method for coal mine disaster risks based on multi-task learning. *Reliab. Eng. Syst. Saf.* **2025**, *260*, 111002. [CrossRef]

https://doi.org/10.3390/fire9070267

27. Vo, D.-T.; Lee, J. Physics-informed Fourier neural operator for the bending analysis of bi-directional functionally graded beams with variable cross-sections. *Eur. J. Mech./A Solids* **2026**, *115*, 105798. [CrossRef]
28. Zoumb, P.A.W.; Wang, M.; Kouame, A.R. Fourier series-based reliability analysis of train-bridge interaction under crosswind action using fuzzy random uncertainty. *Structures* **2025**, *81*, 110148. [CrossRef]
29. Li, X.; Yang, H. Prediction of gas emission zone width in laneway's coal seam based on slippage effect of gas seepage. *Energy Sources Part A Recovery Util. Environ. Eff.* **2020**, *46*, 8821–8837. [CrossRef]
30. Lu, C.; Li, S.; Xu, N.; Qin, Y. A knowledge-data-structure hybrid-driven multi-characteristic fusion weighting model—Application of coal and gas outburst risk assessment. *Expert Syst. Appl.* **2026**, *299*, 130201. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

https://doi.org/10.3390/fire9070267

MDPI AG
Grosspeteranlage 5
4052 Basel
Switzerland
Tel.: +41 61 683 77 34

MDPI Books Editorial Office
E-mail: books@mdpi.com
www.mdpi.com/books



Disclaimer/Publisher's Note: The title and front matter of this reprint are at the discretion of the Topic Editors. The publisher is not responsible for their content or any associated concerns. The statements, opinions and data contained in all individual articles are solely those of the individual Editors and contributors and not of MDPI. MDPI disclaims responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.





Academic Open
Access Publishing

mdpi.com

ISBN 978-3-7258-9080-4