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 I

Edited by
Moulay A. Akhloufi and Mozhdeh Shahbazi

mdpi.com/topics

MDPI

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

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

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 I
ISBN 978-3-7258-9077-4 (Hbk)
ISBN 978-3-7258-9078-1 (PDF)
https://doi.org/10.3390/books978-3-7258-9078-1

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**

Walter Garcia-Suabita, Mario José Pacheco and Dolors Armenteras
Multitemporal Dynamics of Fuels in Forest Systems Present in the Colombian Orinoco River Basin Forests
Reprinted from: *Fire* **2024**, *7*, 171, https://doi.org/10.3390/fire7060171 . . . . . **1**

Qiuping Yu, Yaqin Zhao, Zixuan Yin and Zhihao Xu
Wildfire Susceptibility Prediction Based on a CA-Based CCNN with Active Learning Optimization
Reprinted from: *Fire* **2024**, *7*, 201, https://doi.org/10.3390/fire7060201 . . . . . **14**

Xiangsheng Li and Yongquan Liang
Fire-RPG: An Urban Fire Detection Network Providing Warnings in Advance
Reprinted from: *Fire* **2024**, *7*, 214, https://doi.org/10.3390/fire7070214 . . . . . **29**

Zhi-Hai Li, An-Chi Shi, Huai-Xian Xiao, Zi-Hao Niu, Nan Jiang, Hai-Bo Li and Yu-Xiang Hu
Robust Landslide Recognition Using UAV Datasets: A Case Study in Baihetan Reservoir
Reprinted from: *Remote Sens.* **2024**, *16*, 2558, https://doi.org/10.3390/rs16142558 . . . . . **51**

Tianjun Qi, Xingmin Meng and Yan Zhao
Landslide Susceptibility Assessment in Active Tectonic Areas Using Machine Learning Algorithms
Reprinted from: *Remote Sens.* **2024**, *16*, 2724, https://doi.org/10.3390/rs16152724 . . . . . **73**

Pappu Kumar Yadav, J. Alex Thomasson, Robert Hardin, Stephen W. Searcy, Ulisses Braga-Neto, Sorin C. Popescu, et al.
AI-Driven Computer Vision Detection of Cotton in Corn Fields Using UAS Remote Sensing Data and Spot-Spray Application
Reprinted from: *Remote Sens.* **2024**, *16*, 2754, https://doi.org/10.3390/rs16152754 . . . . . **89**

Qiong Wu, Yi-Xuan Shou, Yong-Guang Zheng, Fei Wu and Chun-Yuan Wang
Detecting Hailstorms in China from FY-4A Satellite with an Ensemble Machine Learning Model
Reprinted from: *Remote Sens.* **2024**, *16*, 3354, https://doi.org/10.3390/rs16183354 . . . . . **111**

Sisi Li, Sheng Hu, Lin Wang, Fanyu Zhang, Ninglian Wang, Songbai Wu, et al.
Quantifying the Geomorphological Susceptibility of the Piping Erosion in Loess Using LiDAR-Derived DEM and Machine Learning Methods
Reprinted from: *Remote Sens.* **2024**, *16*, 4203, https://doi.org/10.3390/rs16224203 . . . . . **135**

Gang Qin, Shixin Wang, Futao Wang, Suju Li, Zhenqing Wang, Jinfeng Zhu, et al.
Flooded Infrastructure Change Detection in Deeply Supervised Networks Based on Multi-Attention-Constrained Multi-Scale Feature Fusion
Reprinted from: *Remote Sens.* **2024**, *16*, 4328, https://doi.org/10.3390/rs16224328 . . . . . **160**

Yankui Li, Wu Zhu, Jing Wu, Ruixuan Zhang, Xueyong Xu and Ye Zhou
DBSANet: A Dual-Branch Semantic Aggregation Network Integrating CNNs and Transformers for Landslide Detection in Remote Sensing Images
Reprinted from: *Remote Sens.* **2025**, *17*, 807, https://doi.org/10.3390/rs17050807 . . . . . **177**

Chenmeng Zhao, Like Zhao, Ka Zhang, Yinghua Ren, Hui Chen and Yehua Sheng
Smoke and Fire-You Only Look Once: A Lightweight Deep Learning Model for Video Smoke and Flame Detection in Natural Scenes
Reprinted from: *Fire* **2025**, *8*, 104, https://doi.org/10.3390/fire8030104 . . . . . **201**

Sudath Loku-Pathirage, Jiye Chen and Min Fu
Predicting Railway Slope Failure Under Heavy Rainfall Using the Soil Moisture Extended Cohesive Damage Element Method
Reprinted from: *GeoHazards* **2025**, *6*, 14, https://doi.org/10.3390/geohazards6010014 . . . . . **217**

Nicholas LaHaye, Anastasija Easley, Kyongsik Yun, Hugo Lee, Erik Linstead, Michael J. Garay and Olga V. Kalashnikova
Development and Application of Self-Supervised Machine Learning for Smoke Plume and Active Fire Identification from the Fire Influence on Regional to Global Environments and Air Quality Datasets
Reprinted from: *Remote Sens.* **2025**, *17*, 1267, https://doi.org/10.3390/rs17071267 . . . . . **236**

Yuxuan Li, Lisha Nie, Fangrong Zhou, Yun Liu, Haoyu Fu, Nan Chen, et al.
Improving Fire and Smoke Detection with You Only Look Once 11 and Multi-Scale Convolutional Attention
Reprinted from: *Fire* **2025**, *8*, 165, https://doi.org/10.3390/fire8050165 . . . . . **261**

Feifan Lu, Guifang Zhang, Tonghao Wang, Yumeng Ye and Qinghao Zhao
Geographically Weighted Random Forest Based on Spatial Factor Optimization for the Assessment of Landslide Susceptibility
Reprinted from: *Remote Sens.* **2025**, *17*, 1608, https://doi.org/10.3390/rs17091608 . . . . . **290**

Rui Zou, Zhihui Xin, Guisheng Liao, Penghui Huang, Rui Wang and Yuhu Qiao
A Fire Segmentation Method with Flame Detail Enhancement U-Net in Multispectral Remote Sensing Images Under Category Imbalance
Reprinted from: *Remote Sens.* **2025**, *17*, 2175, https://doi.org/10.3390/rs17132175 . . . . . **318**

Haipeng Sun and Tao Yao
Multi-Scale Construction Site Fire Detection Algorithm with Integrated Attention Mechanism
Reprinted from: *Fire* **2025**, *8*, 257, https://doi.org/10.3390/fire8070257 . . . . . **343**

Omar Elfahim, Kokou Laris Edjinedja, Johan Cossus, Mohamed Youssfi, Oussama Barakat and Thibaut Desmettre
A Systematic Literature Review of Artificial Intelligence in Prehospital Emergency Care
Reprinted from: *Big Data Cogn. Comput.* **2025**, *9*, 219, https://doi.org/10.3390/bdcc9090219 . . . **365**

Jiacheng Yuan and Baiqing Sun
Multi- Objective Optimization Model for Emergency Evacuation Based on Adaptive Ant Colony Algorithm
Reprinted from: *AI* **2025**, *6*, 203, https://doi.org/10.3390/ai6090203 . . . . . **398**

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

fire

MDPI

Article

Multitemporal Dynamics of Fuels in Forest Systems Present in the Colombian Orinoco River Basin Forests

Walter Garcia-Suabita [1,*], Mario José Pacheco [2] and Dolors Armenteras [1]

[1] Grupo de Investigación en Ecología del Paisaje y Modelación de Ecosistemas, Departamento de Biología, Facultad de Ciencias, Universidad Nacional de Colombia, Bogotá 111321, Colombia
[2] Departamento Estadística, Universidad el Bosque, Bogotá 111321, Colombia
* Correspondence: walter.suabita@gmail.com

Abstract: In Colombia's Orinoco, wildfires have a profound impact on ecosystem dynamics, particularly affecting savannas and forest–savanna transitions. Human activities have disrupted the natural fire regime, leading to increased wildfire frequency due to changes in land use, deforestation, and climate change. Despite extensive research on fire monitoring and prediction, the quantification of fuel accumulation, a critical factor in fire incidence, remains inadequately explored. This study addresses this gap by quantifying dead organic material (detritus) accumulation and identifying influencing factors. Using Brown transects across forests with varying fire intensities, we assessed fuel loads and characterized variables related to detritus accumulation over time. Employing factor analysis, principal components analysis, and a generalized linear mixed model, we determined the effects of various factors. Our findings reveal significant variations in biomass accumulation patterns influenced by factors such as thickness, wet and dry mass, density, gravity, porosity, and moisture content. Additionally, a decrease in fuel load over time was attributed to increased precipitation from three La Niña events. These insights enable more accurate fire predictions and inform targeted forest management strategies for fire prevention and mitigation, thereby enhancing our understanding of fire ecology in the Orinoco basin and guiding effective conservation practices.

Keywords: Orinoco; wildfires; fuel accumulation; ecosystem dynamics; La Niña

1. Introduction

The Colombian Orinoquía, spanning an area of 1,080,000 km^2, is divided between Venezuela (71%) and Colombia (29%) [1]. Dominated by gallery forest and savannas [2], fire as a natural disturbance has been documented to spread well in herbaceous vegetation, occasionally entering the standing forests [3,4]. This region plays a vital role in Colombia's agriculture and forestry sectors, serving as a crucial source of rice, soybeans, corn, beans, biofuels, and timber [5]. However, the accelerated land use changes derived from activities such as mining, illicit crop cultivation, and agriculture have led to significant alterations in land cover and land use, thereby impacting the region's fire regimes. These changes are particularly consequential for the savannas, which play a key role in regulating water and carbon dynamics. Through the tree–grass dynamic, savannas contribute to organic matter storage and facilitate seasonal burns that mitigate CO_2 accumulation in the atmosphere, curbing greenhouse gas increases and the resultant temperature rises [6,7].

Gallery forests, another ecosystem affected by fire-related issues, have evolved a symbiotic relationship with fires in adjacent savannas, developing differential functional responses and resilience to these events [1,8]. However, the expansion of agricultural frontiers and deforestation driven by climate change and logging have altered fire patterns. This alteration is partly due to increased wind speeds, resulting from the removal of trees that previously functioned as natural wind barriers [9]. Consequently, there has been a loss of biodiversity and soil degradation, reducing soil moisture and intensifying fires [10].

These ecological changes adversely affect human populations, leading to public health concerns about diminished air quality [11].

Understanding fire dynamics fundamentally involves the role of fuels, defined as biological materials that facilitate the propagation and dispersion of fire in natural systems. The combustion and emissions from forest fires are significantly influenced by the type, quantity, characteristics, and condition of these wildland fuels. Despite the development of various methodologies, accurately quantifying these factors remains challenging due to complex variables like vegetation type, moisture, and terrain slope [12]. Fuels are critical in determining fire behavior through three main components: ignition (how easily they catch fire), combustibility (how readily they burn), and sustainability (how long they burn) [13]. Fuels are classified based on their shape, size, arrangement, and morphology, with morphology offering the best insight into fire behavior [14]. They are classified as fine and coarse fuels, the former comprising leaves, bark, and branches with a diameter of less than 6 mm and fine stems with a thickness of less than 3 mm. This combustible material is consumed quickly and is easily ignited [15]. Coarse fuels, on the other hand, correspond to branches and trunks with diameters greater than 6 mm. The time it takes for them to burn allows them to be divided into 10 h and 100 h, constituting fuels with a delayed response to fire [15].

The measurement of fuel load presents significant challenges across different ecosystem types, with notable differences between forest systems and open areas, such as savannas, which are the predominant systems in the Orinoco basin. The complexity in measurement arises from the diverse and structurally variable vegetation, particularly native grasses, complicating accurate fuel load estimation [16]. Precise fuel load measurements are crucial for developing various management strategies for these ecosystems, such as the controlled burning of biological material and minimizing escaped fire incidents [17].

This research focused on understanding fuel load characteristics and identifying variables influencing fire occurrence and persistence. We employed statistical methods such as K prototypes, factor analysis of mixed data, and generalized linear mixed models to analyze multitemporal changes in fuel accumulation. These methods were selected for their efficacy in handling mixed data types, offering a detailed insight into fuel load patterns [18]. We hypothesized that fuel load varies over time and is influenced by the morphological and microclimatic conditions of each area under study. This research's goal was to aid in the future development of targeted fuel management practices, aiming to reduce both the frequency and intensity of forest fires in the Orinoco by understanding the dynamics of fuel load variation and its predictors.

2. Materials and Methods

2.1. Area of Study

Within Colombia, the Orinoquia region is the largest biogeographical region in terms of extension, covering 310,000 km^2. Its origin dates back to the Quaternary period after a geological, geomorphological, and edaphic evolution, resulting from a structural depression that generated a geosyncline located between the Andes mountains, the coast, and the Guiana Shield [19]. Its ecological importance lies in its role as a zone of exchange of diversity between the Amazon and the Andes [20]. Within this zone, two ecosystems primarily stand out: gallery forests and savannas, with the latter being the ecosystems with the greatest dominance and importance at the structural level due to their extension and resilience to the periodic fires that generate changes in the composition of species [21].

Our study was conducted in the Orinoco River basin, in two Departments, Vichada and Arauca, most of the evaluated forests are secondary, in which their composition varies according to their disturbance, with burned forests presenting a lower species composition with 61 species and 37 genera, while unburned forests are represented by 59 and 77 species [4], respectively. Regarding their structure, systems without burns have an average basal area of 3.0 m^2/0.1 ha ($\pm$1.31), a height of 16 m ($\pm$1.81), and a density of 613.3 density/ha ($\pm$129.77); for burned areas, the basal area is 1.68 m^2/0.1 ha ($\pm$0.60), a

height of 13.4 m (±3.31), and a density of 271.9 (±164.76), with sampling plots and transects established in the evergreen floodplain forests. In Vichada, the study was conducted in three civil society natural reserves: Bonjonawi, comprising 4680 ha; Los Robles, with 155.68 ha; and Doña Ana, with 1200 ha. We also included sampling plots at the National Natural Park, a protected area covering 548,000 ha, and a forest plantation (Forest First) composed of stands of *Eucalyptus pellita* and *Acacia mangium* whose purpose is the acquisition of wood and paper pulp, primarily by the Forest First reforestation company; Forest First encompasses 31,758 ha. Each system has been affected by fires to varying degrees and comprises natural systems such as gallery forests, savannas, and Moriche (*Mauritia flexuosa)* palm swamps (Figure 1a). This area experiences an average temperature of 26.9 °C and receives an average annual precipitation of 2366 mm. The predominant soils are acidic oxisols with low water retention capacity [22].

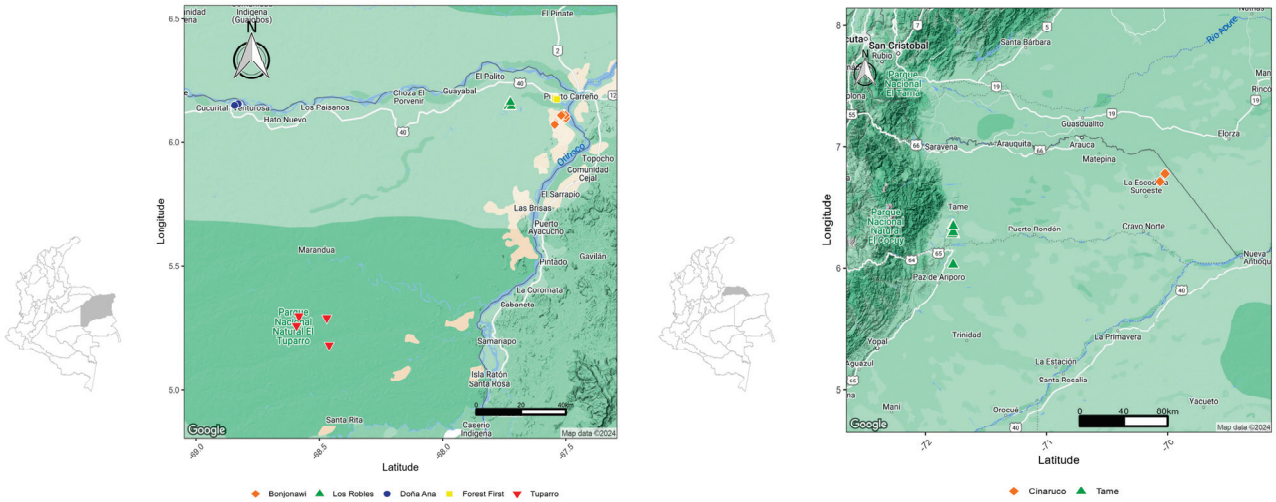


Figure 1. Collection sites for dead fuel (**left**) Vichada, (**right**) Arauca.

Additionally, sampling efforts extended to the Department of Arauca in the Cinaruco Management District spanning 331,448 ha, and to a group of farms (Altamira and La Fortuna) in the municipality of Tame, totaling 367 hectares. This zone is characterized by a variety of natural systems including savannas, gallery forests, where Moriche (*Mauritia flexuosa* L. f.) palm swamps, Saladillo (*Caraipa llanorum* Cuatrec) marshes, and Congrio (*Acosmium niten* Vogel) palm forests dominate, and the area has an average annual temperature of 32.3 °C and an annual precipitation of 1881 mm [23] (Figure 1b). The savannas feature herbaceous species, dicotyledons (e.g., *Polygala* spp., *Chamaecrista* spp., and *Drosera sessilifloria*), and trees (e.g., *Byrsonima* spp., *Bowdichia virgilioides*, *Curatella americana).* Gallery forests, on the other hand, are mainly composed of species from the Lauraceae, Chrysobalanaceae, and Melastomataceae families [24]. For both departments, unimodal behaviors are presented in which there are two rainy periods corresponding to the wet and dry seasons, with the months from January to March having the least precipitation for Vichada and Arauca, and the periods from May to July having the highest amount of rainfall, with the dry season being when the highest number of fires occur in the region [25].

2.2. Categorization of Fire Intensity and Severity

The degree of impact on each forest differed, so three conditions were defined based on the traces left in forests with diameters greater than 10 (DAP > 10), using parameters proposed by Keeley [26] and Pausas and Keeley [27] for changes in vegetation after a fire. The proposed classification includes trees with diameters greater than 10 that have burned crowns, completely burned detritus, and a canopy opening greater than 70%. This condition is classified as a burned forest of high severity and intensity (BQ). Finally, non-

burned forests (BNQ) were considered; these are natural areas where there have been no fires, logging, or grazing in the last 30 years [23].

The selection of samples for the measurement of fuel load was made for the three forest conditions, establishing six (6) subplots of 0.1 ha, with a horizontal area of 10 m, at a vertical distance of 100 m, with up to a total of 18 plots, specifically for the department of Vichada; in Arauca, we examined only two types of conditions: burned forest and unburned forest. For each experimental unit, it was guaranteed that they were separated by 400 m. Fuel measurements were conducted in the years 2017, 2019, 2021, and 2023, and a total of 5426 transects were evaluated, with 5012 for Vichada, corresponding to 3944 for burned areas and 1068 for unburned areas, while for Arauca, 404 were taken into account, distributed in 202 unburned zones and 212 burned zones. The areas experienced fires in different years. In Bonjonawi and Doña Ana, the fires occurred in 2015, while in Los Robles, it occurred in 2018. On the other hand, in Tuparro and Forest First, the fires affected the properties in 2021. In the case of Tame, the burned areas were recorded in 2019, and in Cinaruco, the fires took place in 2018.

2.3. Woody Detritus Sampling

For the measurement of dead material, the previously established plots were used and two 20 m Brown linear transects were constructed [15]. In each of them, two variables were taken for the dead woody fuel: fine woody detritus (FWD) and coarse woody detritus (CWD). These detritus were classified according to their size class, considering their diameter and the time it takes to burn 0–0.6 cm (1 h to burn), 0.6–2.5 cm (10 h), 2.5–8 cm (100 h), and 8 cm (1000 h)

For the collection of dead biomass, samples were taken from the first 2 m of fuels with ignition times of 1 and 10 h. Subsequently, detritus with ignition times of 100 h were collected from the next five meters, and fuels with ignition times of 1000 h were collected along the entire Brown transect. For each sample, a 3 cm thick cylindrical wood segment was obtained, and measurements of wet weight, dry weight, and density were conducted following the method proposed in [28]. Additionally, for wood pieces with diameters ranging from 2.5 cm to 8 cm, the azimuth, inclination, weight, thickness, and hardness (penetration) were determined. The values for the penetration variable were obtained using a penetrometer, with 20 blows applied to the piece under a constant force of 400 kg. This allowed the classification of detritus as either healthy or rotten, with the latter being those that did not withstand more than 20 blows, after the collection of the debris, they were classified according to their shape, type, and morphology.

2.4. Data Analysis

A total of 26 variables were evaluated, including the zone, indicating the specific locations of fuel load measurement; year, representing the period of measurement; condition, denoting the category of fire events; plot, which signifies the experimental sampling unit; and subplot, delineating subgroups within the experimental unit. Additionally, there were categorical variables like diameter class (cm) for the diameter class of the detritus; fuel class (h), indicating the combustion time of the fuel; and shape, representing the shape of the detritus. Other variables, specifically those of a numerical nature, included diameter detritus (cm) for the diameter of the detritus; azimuth degrees (°) representing the degree of detritus position; inclination (°) indicating the position of the detritus; and others like wet mass (g) for the wet weight of the detritus, volume (m^3) denoting the volume of the detritus, and fuel load (kg/m^2) representing the fuel load of the detritus. These variables collectively provided a comprehensive understanding of the fuel load characteristics in the study area and were crucial for analyzing the dynamics and variations in fuel accumulation over time. For the above, 5426 observations were considered.

The analyses presented were entirely developed using R version 4.2.3. Prior to the data analysis, data imputation was performed using the MICE algorithm, using the predictive mean matching (PMM) on numerical variables, which was performed due to the high

percentage of missing values and to obtain a precise estimation of patterns and changes in load fuel. A multiple imputation was conducted with a maximum of 10 interactions [29]. Subsequently, boxplots and a Kruskal–Wallis test were constructed to identify differences and changes in fuel load based on the different categorical variables collected during the estimation of this variable.

To identify groups among all variables, a unsupervised technique k-prototypes similar to k-means determined groups in the adjacent data, where the analysis was conducted using the Gower distance, taking into account the mixed types of categorical and continuous variables. Additionally, to determine the number of clusters, we implemented the method of Mclaine and Silhouette, and then, using the elbow criterion, the optimal number of groups was selected using the mixed data taken in the forests with different degrees of fire intensity [30]. These procedures were performed using the clustMixType package version 1.0 [31].

To detect trends and relationships between the proposed variables to explain fuel load, a factor analysis of mixed data extension with MCA (multiple correspondence analysis) and generalization PCA (principal components analysis) was performed, implemented in the Factomine R package [32]. Finally, to determine the effect of the variables recorded on the fuel load, a generalized linear mixed model (GLMM). These are versatile statistical tools that excel in various modeling scenarios. First and foremost, they are particularly suitable for data with a hierarchical or correlated structure, such as those involving repeated measurements over time or data collected from multiple sites. In such cases, GLMMs can capture the structure more accurately than generalized linear models (GLMs), allowing for a more precise analysis that is better suited to the nature of the data [33]. Model adjustment to an inverse Gaussian distribution was performed using the lme4 package [34]. After verification using a likelihood ratio test [35], a model that showed significance variables was reported, and the marginal and conditional R-squared was calculated using the Mumin package [36]. Additionally, the climatic values of precipitation and temperature for each year of fuel measurement were obtained from information collected through the Meteostat website.

3. Results

The fuel load found in each of the evaluated areas indicated that the highest amount for this component was found in the Tuparro National Natural Park with 24.9 ($\pm$0.24) tons per hectare, followed by Tame with 22.8 ($\pm$5.96), Bonjonawi 19.8 ($\pm$1.31), Cinaruco 17.7 ($\pm$4.26), followed by Los Robles 12.3 ($\pm$2.24), Forest First 11.9 ($\pm$3.24), and Doña Ana with 11.1 ($\pm$1.30), respectively. Within the most common classes, fine fuels accounted for 87.5% of the total evaluated, while coarse fuels represented 12.5%. Likewise, the most frequently observed shape in the transects was cylindrical, with 5353 records. Finally, the debris most present in the evaluated transects were those whose ignition took at least 10 h. The reported fuel values specifically for areas with water tributaries such as the Arauca and Orinoco rivers reached an average value of 20.5 tons per hectare ($\pm$0.19), while the areas adjacent to the Vichada, Bita, and Orinoco rivers presented a fuel load of 17.4 ($\pm$0.01) tons per hectare.

Our comparative analysis revealed significant differences in median fuel loads across the study areas ($\chi^2(1) = 21.82$, $p \leq 0.0001$, $n = 5426$), with Vichada department demonstrating notably higher levels. This variation was due to a more uniform distribution of fuel in Vichada, leading to greater measurement variability (Figure 2a). The zones exhibited significant differences ($\chi^2(6) = 302.5$, $p \leq 0.0001$, $n = 5426$); specifically, Bonjonawi area in Vichada department recorded the highest fuel load, accounting for 50% of the records, surpassing those in Tame and Cinaruco. The data from Bojonawi and Cinaruco exhibited a similar pattern, whereas Tame showed a more constricted distribution, highlighting significant variability and outliers indicating high fuel load (Figure 2b).

Regarding the conditions assessed—unburned and burned forests—a pronounced variation was observed, particularly due to the distribution narrowness (Figure 2c). Over the evaluation period, median fuel loads showed consistency for the initial two years, while the values for 2022 and 2023 were notably lower. The distribution of fuel load exhibited

a strong trend of extreme data for 2017, 2022, and 2023, while 2019 showed a positive skewness behavior, with constant values and slight variation (Figure 2d).

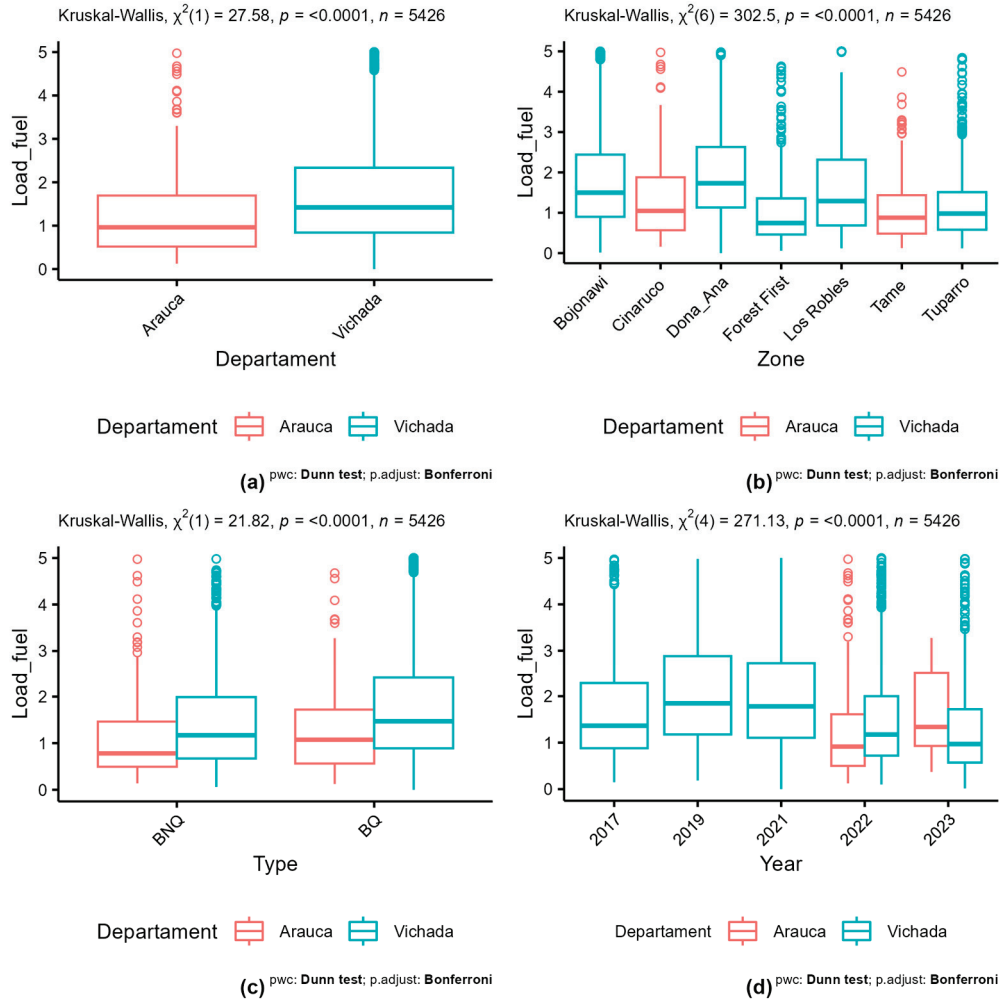


Figure 2. Boxplot of fuel load showing four categories of variables, (**a**) department, (**b**) zone, (**c**) condition, and (**d**) year.

A cluster analysis based on the sum of squares indicated an optimal number of two clusters, as determined by the elbow criterion, with a Maclain value of −0.23 and silhouette value of 0.12. (Figure 3a). When using specific gravity as the reference variable, the analysis revealed distinct separation between the clusters, with one group characterized by lower fuel loads and specific gravity of detritus (Figure 3b). Regarding distance, one cluster was notably associated with higher fuel loads and greater distances. This analysis also highlighted variations in the debris arrangement across transects (Figure 3c). Lastly, regarding porosity, the highest clustering was observed towards high porosity values; in contrast, there was a group with fewer records with low porosity values and fuel loads (Figure 3d).

The factor analysis of mixed data revealed that the first two components accounted for 23.2% of the variation, with 14.3% attributed to the first component and 8.9% to the second, with no relevant information about variation. This led to the selection of two components based on their ability to retain the most significant information regarding fuel load variables (Figure 4a). A quantitative analysis showed a strong correlation between weight-related variables and fuel loads, whereas porosity demonstrated an inverse relationship with dasometric variables and fuel load, alongside a notable correlation with density. Some variables, however, poorly explained the variation, indicating potential exclusion from

further analysis (Figure 4b). Categorical variable analysis identified two distinct groups, differentiated by the shape, type of detritus, and distance, as well as by sampling zones (Figure 4c).

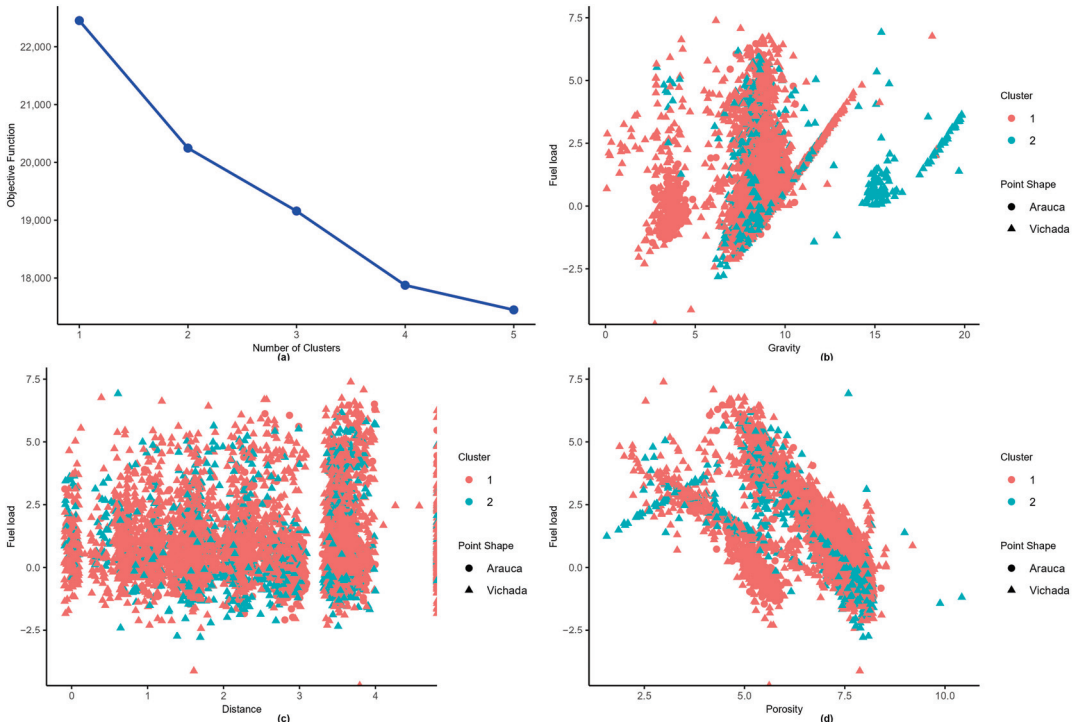


Figure 3. K prototypes of the class finding with all variables, (**a**) number of clusters, (**b**) groups to fuel load and diameter, (**c**) groups to fuel load and distance, and (**d**) groups to fuel load and porosity.

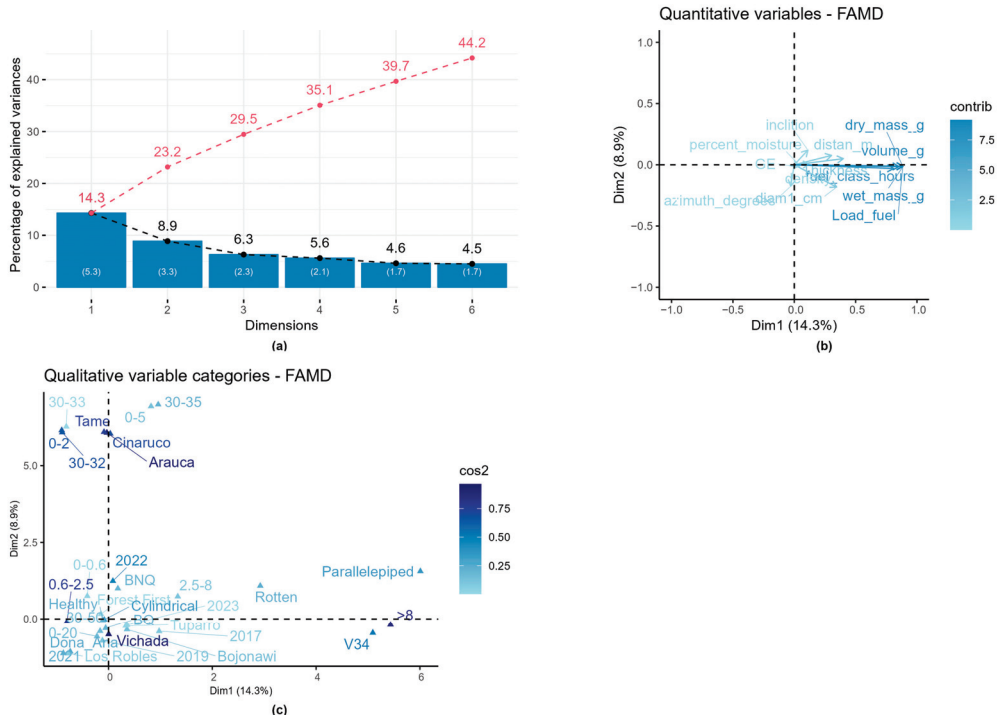


Figure 4. FAMD—Factor analysis of mixed data, (**a**) proportion of variance explained by each dimension, the dashed line red represents the variation explained by each component and the dashed black line eigen values, (**b**) biplot quantitative variables, (**c**) biplot qualitative variables.

The generalized linear mixed model (GLMM) analysis, utilizing an inverse Gaussian distribution, provided insights into fuel load variations in the Orinoco River basin, specifically within Vichada. The model demonstrated a marginal R-squared of 59.4% and a conditional R-squared of 64.4%, with an Akaike information criterion (AIC) of 23,396, Bayesian information criterion (BIC) of 23,471, and a log-likelihood of −11,686. Within the model, variables such as precipitation, percent moisture, and porosity had a significant negative effect on fuel loads. In contrast, thickness, diameter, and azimuth degrees had a significant positive effect on the response variable (Table 1).

Table 1. Results of generalized linear mixed model (GLMM) for fuel load.

Variable	β	Std. Error	*p*
Thickness	0.565	0.399	5×10^{-5}
Precipitation	−0.152	0.041	0.00244
Diameter	0.425	0.032	2×10^{-16}
Porosity	−0.401	0.014	2×10^{-16}
Percent moisture	−0.086	0.009	2×10^{-16}
Azimuth degrees	0.053	0.025	0.035
Inclination	0.542	0.064	2×10^{-16}
Constant	2.542	0.176	2×10^{-16}

In the model, the variables precipitation, porosity, and percent moisture, given an increase of one unit, tended to decrease the fuel load by 0.152, 0.401, and 0.086, respectively, in terms of kilograms per hectare. In contrast, the variables thickness, azimuth degrees, inclination, and diameter, given an increase of one unit, tended to increase the fuel load by 0.565, 0.053, 0.542, and 0.425 kg per hectare, respectively. As for the intercept, it was observed that when the variables had zero values (for interpretation purposes), they tended to take a value of 2.542 tons per hectare.

$$\begin{gathered} yij \sim \text{Inverse Gaussian}(\mu ij) \\ log(\mu ij) = \beta 0 + \beta 1 \times 1ij - Precipitation + \beta 2 \times 2ij - Percent\ moisture + \beta 3 \times 3ij - \\ Thickness + \beta 4 \times 4ij - Porosity + \beta 5 \times 5ij - Inclination + \beta 6 \times 6ij - Diameter + \beta 7 \times \\ 7ij - Azimuth\ degrees + b0i + b1j \end{gathered} \tag{1}$$

The mixed-effects generalized model, employing an inverse Gaussian distribution for Arauca, yielded a marginal R-squared of 76.4% and a conditional R-squared of 76.8%, effectively elucidating the variation in fuel loads within systems situated at the base of the Orinoco River. Furthermore, the model exhibited favorable values for the Akaike criterion (1891), Bayesian criterion (1927.3), and likelihood (−936.6). Notably, variables such as thickness and diameter exerted a positive influence on fuel loads, while the percent moisture and porosity demonstrated a significant negative impact on the response variable (Table 2).

Table 2. Results of generalized linear mixed model (GLMM) for fuel load in Arauca.

Variable	β	Std. Error	*p*
Thickness	0.568	0.13256	1.83×10^{-5}
Diameter	0.335	0.129	2×10^{-16}
Porosity	−0.922	0.054	2×10^{-16}
Percent moisture	−0.28	0.018	2×10^{-16}
Constant	1.481	0.145	2×10^{-16}

In the model, when the variables porosity and percent moisture were increase by one unit, the fuel load tended to decrease by 0.922 and 0.28 kg per hectare, respectively. Conversely, when the variables thickness and diameter were increased by one unit, the fuel load tended to increase by 0.568 and 0.335 kg per hectare, respectively. Regarding the intercept, it was observed that when the variables had zero values (for interpretation purposes), it tended to be 1.481 tons per hectare.

$$\begin{gathered} yij \sim Inverse\ Gaussian(\mu ij) \\ log(\mu ij) = \beta 0 + \beta 1 \times 1ij - Thickness + \beta 2 \times 2ij - Diameter + \beta 3 \times 3ij - \\ Percent\ moisture + \beta 4 \times 4ij - Porosity + b0i + b1j \end{gathered} \quad (2)$$

4. Discussion

This study presents a comprehensive analysis of multitemporal changes in fuel accumulation in forest systems within the Colombian Orinoco River basin. Our findings revealed significant variations in fuel load across different departments, zones, conditions, and years, with notable implications for forest management and conservation strategies in this region. Our results indicated a higher median fuel load ($\chi^2(1) = 21.82$, $p \leq 0.0001$, $n = 5426$) in the Vichada department compared to the others, with the Bonjonawi area exhibiting the highest fuel loads. This variability can be attributed to several factors, including differences in vegetation types, forest management practices, and climatic conditions. Understanding these differences is crucial for developing targeted strategies for fire risk management and conservation. Although they are used in forest plantations, the use of silvicultural practices such as pruning and thinning in natural forests, as a management tool for fuel, is a feasible and practical alternative that can reduce the incidence of fires [37]. However, certain limitations have to be considered, since these treatments are conducted in homogeneous forests with dominant species such as pines and eucalyptus [38]. Additionally, in this study, we did not directly measure parameters such as fuel bed depth, and fuel moisture content within the examined ecosystems. Future research could benefit from incorporating these variables to provide a more comprehensive understanding of fire dynamics across different ecological settings.

The use of an inverse Gaussian distribution in our mixed-effects model was instrumental in capturing the skewness and kurtosis typical of ecological data, particularly in fuel load measurements. This approach provided a more accurate representation of the underlying data distribution, enhancing the reliability of our findings. Furthermore, the multiple correspondence analysis highlighted the complex interactions between numerous factors influencing fuel loads. The decision to potentially remove certain variables from the analysis was based on their low explanatory power, which suggests that they might not significantly contribute to fuel load variations. Many studies related to fuel load have made inferences according to normal or gamma-adjusted distributions [39–41]. Although these assumptions are adequate with an optimal number of samples, they justify the results very well. However, the use of an appropriate distribution, as in this case, allows for improved prediction and avoidance of erroneous interpretations regarding type one and two errors in the information on detritus quantity.

The decrease in fuel load over the past few years may be attributed to climate variability events, including three La Niña events in the years 2017, 2020, and 2021. Their impact was associated with increased precipitation, a factor that served as one of the main drivers for the reduction e in fuel load. This phenomenon was evident in the Orinoco River basin, due to the fluctuation of water levels causing a displacement of debris, which was evident in Doña Ana and Bonjonawi, areas that reached an average distance of 1001 km. This proximity conditioned the temporary quantity of debris, in addition to increasing decomposition rates due to increased humidity and consequent biological activity. This highlighted a self-management system of fuels by natural systems in the Vichada department [42].

The presence of outliers, especially in areas like Tame, raises important questions about the factors contributing to these anomalies. These could range from microclimatic conditions to specific forest management practices. Future studies should aim to investigate these outliers in more detail, to understand their causes and implications. One of the main factors that has attracted interest and generated anomalous behaviors in fuel estimation is climate change, which has induced a reduction in the humidity of detritus [43]. This is of interest because forest fires are one of the main causes of air pollution, and changes in climatic conditions can affect this behavior, generating health problems [44].

Our results suggest an increase in fuel load across the transect distance, which is associated with edge-type behaviors, where the transition between forests and savannas plays a functional role in the dynamics and formation of open systems with low vegetation cover. Taking this into account, although savannas experience the highest frequency of fires, forests present a higher amount of biomass, due to their greater production of plant material by trees, as well as the effect of the functional diversity of these systems, specifically in the production of resins and the flammability of these materials, as reported by Valencia et al. [3]. On the other hand, as highlighted by Armenteras et al. [4], there is lower species diversity at the edge, caused by the frequency of fires at the forest fringe, which limits the growth of species stems and biomass generation by trees located in this segment.

In future work, the adaptation and selection of machine learning models for land cover and risk classification could efficiently determine the amount of fuel in an area, as well as the speed of fire spread and its impact on natural systems. Additionally, it is essential to integrate different environmental variables to improve model accuracy. Among these variables, precipitation, wind speed, and temperature stand out as sources of variation. This integration will enable the development of dynamic rather than static models, enhancing fire prediction based on fuel load [45,46].

5. Conclusions

This study provided valuable insights into the dynamics of fuel accumulation in the Colombian Orinoco River basin's forest systems. Fuel load variability over time is influenced by several factors, primarily the morphological characteristics of detritus such as density and thickness. These variables tend to affect the accumulation of fuels, which in turn conditions the speed of propagation and ignition of the fuels. Other characteristics of detritus include the amount of moisture that it can retain, which is specifically associated with the speed and ability to spread in forests. Therefore, the monitoring of environmental variables can be one of the main tools for the management and control of fires in natural systems. While the current management of fuels is specific to plantations, strategies such as prescribed burning or manual removal of dead biomass may be eventually implemented in natural systems. These strategies have demonstrated a reduction in the frequency of fires and an improvement in soil biodiversity. Therefore, correct implementation of these strategies would enhance fire management in the studied areas, promoting conservation of these natural areas and reducing the impact on the diversity of these forests.

The observed variations in fuel load underscore the need for region-specific management strategies, reflecting the unique ecological characteristics of each area. The insights gained from this study are vital for informing forest management practices in the Colombian Orinoco River basin. By understanding the spatial and temporal patterns of fuel accumulation, forest managers can better predict areas at higher risk of wildfires and implement more effective fire prevention and control measures. Additionally, our findings can guide reforestation and conservation efforts, ensuring that they are tailored to the specific conditions of each area. Our findings can contribute to a deeper understanding of these systems, offering a foundation for future research and practical applications in forest management and conservation.

These insights enable more accurate fire risk predictions and can inform targeted forest management strategies for fire prevention and mitigation. This research enhances our understanding of fire ecology in the Orinoco basin, guiding effective conservation and potential fuel management practices.

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

Funding: The data analyzed in this paper were collected over the years with funding provided by the United States Agency for International Development—Partnership for Enhanced Engagement in Research (PEER) and the US National Academies of Science (Subaward 2000007526PEER Cycle 5) through the project 'Degradation of Tropical Forest in Colombia: Impacts of Fire'. Additional support came from the Colombian Administrative Department of Science, Technology and Innovation (COLCIENCIAS) project No 110180863738 (CT-247-2019) and the project 'Participatory Design of Strategies for Reducing Forest Fires, Conserving Biodiversity, and Regional Development in Multifunctional Landscapes of Vichada (Ministry of Science and the General System of Royalties of Colombia, BPIN 2020000100456), and finally the project funded by Universidad Nacional de Colombia Ecological and Social Characteristics that Influence the Dynamics of Forest Fires in the Cinaruco National Integrated Management District all awarded to DAP.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All of the data for this study have been archived at the Universidad Nacional de Colombia. Bogotá, Colombia.

Acknowledgments: We would like to thank the Ecolmod team, particularly Maria Meza and Alejandra Reyes, as well as Jose Luis Camargo and Nathalia Moreno, who processed part of the material collected in the latest campaigns.

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

References

1. Gassón, R.A. Orinoquia: The archaeology of the Orinoco River Basin. *J. World Prehist.* **2002**, *16*, 237–311. [CrossRef]
2. Rangel-Ch, J.O.; Sánchez-C, H.; Lowy-C, P.; Aguilar-P, M.; Castillo-G, A. *Región de la Orinoquia. Colombia, Diversidad Biótica I*; Instituto de Ciencias Naturales, Convenio Inderena-Universidad Nacional de Colombia: Bogotá, Colombia, 1995; pp. 239–254.
3. Valencia, S.; Salazar, J.F.; Hoyos, N.; Armenteras, D.; Villegas, J.C. Current Forest–Savanna Transition in Northern South America Departs from Typical Climatic Thresholds. *Ecosystem* **2024**, *27*, 61–76. [CrossRef]
4. Armenteras, D.; Meza, M.C.; González, T.M.; Oliveras, I.; Balch, J.K.; Retana, J. Fire threatens the diversity and structure of tropical gallery forests. *Ecosphere* **2021**, *12*, e03347. [CrossRef]
5. Lavelle, P.; Rodríguez, N.; Arguello, O.; Bernal, J.; Botero, C.; Chaparro, P.; Fonte, S.J. Soil ecosystem services and land use in the rapidly changing Orinoco River Basin of Colombia. *Agr. Ecosyst. Environ.* **2014**, *185*, 106–117. [CrossRef]
6. Armenteras, D.; Cabrera, E.; Rodríguez, N.; Retana, J. National and regional determinants of tropical deforestation in Colombia. *Reg. Environ. Change* **2013**, *13*, 1181–1193. [CrossRef]
7. Romero-Ruiz, M.; Etter, A.; Sarmiento, A.; Tansey, K. Spatial and temporal variability of fires in relation to ecosystems, land tenure and rainfall in savannas of northern South America. *Glob. Change Biol.* **2010**, *16*, 2013–2023. [CrossRef]
8. Hoffmann, W.A.; Geiger, E.L.; Gotsch, S.G.; Rossatto, D.R.; Silva, L.C.; Lau, O.L.; Franco, A.C. Ecological thresholds at the savanna-forest boundary: How plant traits, resources and fire govern the distribution of tropical biomes. *Ecol. Lett.* **2012**, *15*, 759–768. [CrossRef] [PubMed]
9. Brando, P.M.; Nepstad, D.C.; Balch, J.K.; Bolker, B.; Christman, M.C.; Coe, M.; Putz, F.E. Fire-induced tree mortality in a neotropical forest: The roles of bark traits, tree size, wood density and fire behavior. *Glob. Change Biol.* **2012**, *18*, 630–641. [CrossRef]
10. Etter, A.; Sarmiento, A.; Romero, M.H. Land use changes (1970–2020) and carbon emissions in the Colombian Llanos. In *Ecosystem Function in Savannas: Measurement and Modeling at Landscape to Global Scales*, 1st ed.; Hill, M.J., Hanan, N.P., Eds.; CRC Press: Boca Raton, FL, USA, 2010; pp. 383–402.
11. Mendez-Espinosa, J.F.; Belalcazar, L.C.; Betancourt, R.M. Regional air quality impact of northern South America biomass burning emissions. *Atmos. Environ.* **2019**, *203*, 131–140. [CrossRef]
12. Weise, D.R.; Wright, C.S. Wildland fire emissions, carbon, and climate: Characterizing wildland fuels. *For. Ecol. Manag.* **2014**, *317*, 26–40. [CrossRef]
13. Gill, A.M.; Zylstra, P. Flammability of Australian forests. *Aust. For.* **2005**, *68*, 87–93. [CrossRef]
14. Pyne, S.J.; Andrews, P.; Laven, R.D. Introduction to wildland fire. In *Fire Management in the United States*; John Wiley & Sons: Hoboken, NJ, USA, 1996.
15. Brown, J.K. *Handbook for Inventorying Surface Fuels and Biomass in the Interior West*; US Department of Agriculture, Forest Service, Intermountain Forest, and Range Experiment Station: Washington, DC, USA, 1981; Volume 129.
16. Russell-Smith, J.; Cook, G.D.; Cooke, P.M.; Edwards, A.C.; Lendrum, M.; Meyer, C.P.; Whitehead, P.J. Managing fire regimes in north Australian savannas: Applying Aboriginal approaches to contemporary global problems. *Front. Ecol. Environ.* **2013**, *11*, 55–63. [CrossRef]
17. Keane, R.E.; Agee, J.K.; Fule, P.; Keeley, J.E.; Key, C.; Kitchen, S.G.; Schulte, L.A. Ecological effects of large fires on US landscapes: Benefit or catastrophe? *Int. J. Wildland Fire* **2008**, *17*, 696–712. [CrossRef]

18. Nelson, K.N.; Turner, M.G.; Romme, W.H.; Tinker, D.B. Landscape variation in tree regeneration and snag fall drive fuel loads in 24-year-old post-fire lodgepole pine forests. *Ecol. Appl.* **2016**, *26*, 2424–2438. [CrossRef]
19. Rangel-Ch, J.O.; Andrade-C, M.G.; Jarro-F, C.; Santos-C, G. (Eds.). *Colombia Diversidad Biótica XX. Territorio Sabanas y Humedales de Arauca (Colombia)*; Universidad Nacional de Colombia-Instituto de Ciencias Naturales, Parques Nacionales Naturales de Colombia: Bogotá, Colombia, 2019; p. 852.
20. Rippstein, G.; Escobar, G.; Motta, F. *Agroecología y Biodiversidad de las Sábanas en los Llanos Orientales de Colombia*; Publicación CIAT no 322; CIAT: Cali, Colombia, 2001; p. 302.
21. Lasso, C.; Usma, J.; Trujillo, F.; Rial, A. *Biodiversidad de la Cuenca del Orinoco: Bases Científicas para la Identificación de áreas Prioritarias para la Conservación y uso Sostenible de la Biodiversidad*; Instituto de Investigación de Recursos Biológicos Alexander von Humboldt: Bogotá, Colombia, 2010; ISBN 978-958-8554-13-6.
22. Pacheco, M.; Peñuela, L.; Solano, C.; Galan, S. *Manejo Forestal Sostenible en Plantaciones en la Cuenca del rio Bita, Vichada, Colombia, Proyecto: "Fortalecimiento Institucional y de Política para Incrementar la Conservación de la Biodiversidad en Predios Privados en Colombia"*; WWF, Fundación Natura, Resnatur: Bogotá, Colombia, 2014; Volume 1.
23. Meza, M.C.; Espelta, J.M.; González, T.M.; Armenteras, D. Fire reduces taxonomic and functional diversity in Neotropical moist seasonally flooded forests. *Perspect. Ecol. Conserv.* **2023**, *21*, 101–111. [CrossRef]
24. Parra, C. Estudio general de la vegetación nativa de Puerto Carreño (Vichada, Colombia). *Caldasia* **2006**, *28*, 165–177.
25. Armenteras-Pascual, D.; Retana-Alumbreros, J.; Molowny-Horas, R.; Roman-Cuesta, R.M.; Gonzalez-Alonso, F.; Morales-Rivas, M. Characterising fire spatial pattern interactions with climate and vegetation in Colombia. *Agric. For. Meteorol.* **2011**, *151*, 279–289. [CrossRef]
26. Keeley, J.E. Fire intensity, fire severity and burn severity: A brief review and suggested usage. *Int. J. Wildland Fire* **2009**, *18*, 116–126. [CrossRef]
27. Pausas, J.G.; Keeley, J.E. Epicormic Resprouting in Fire-Prone Ecosystems. *Trends Plant Sci.* **2017**, *22*, 1008–1015. [CrossRef]
28. Lutes, D.C.; Keane, R.E.; Caratti, J.F.; Key, C.H.; Benson, N.C.; Sutherland, S.; Gangi, L.J. *Fuel Load (FL) Sampling Method*; Forest Service General Technology Report; USDA Forest Service: Washington, DC, USA, 2006; 27p.
29. Van Buuren, S.; Groothuis-Oudshoorn, K. mice: Multivariate imputation by chained equations in R. *J. Stat. Softw.* **2011**, *45*, 1–67. [CrossRef]
30. Huang, Z. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Min. Knowl. Discov.* **1998**, *2*, 283–304. [CrossRef]
31. Szepannek, G.; Aschenbruck, R.; Szepannek, M.G. clustMixType: Clustering Mixed-Type Data in R. R Package Version 1.0. Available online: https://cran.r-project.org/package=clustMixType (accessed on 14 January 2024).
32. Pagès, J. Analyse factorielle de donnees mixtes: Principe et exemple d'application. *Rev. Stat. Appl.* **2004**, *52*, 93–111.
33. Bates, D.; Maechler, M.; Bolker, B.; Walker, S.; Christensen, R.H.B.; Singmann, H.; Green, P. Package 'lme4'. 2009. Available online: http://lme4.r-forge.r-project.org (accessed on 17 December 2023).
34. Correa Morales, J.C.; Salazar Uribe, J.C. *Introducción a los Modelos Mixtos*; Universidad Nacional de Colombia: Bogotá, Colombia, 2016; pp. 30–50.
35. Davis, A.S. Use of the likelihood ratio test on the inverse Gaussian distribution. *Amer. Stat.* **1980**, *34*, 108–110. [CrossRef]
36. Barton, K.; Barton, M.K. MuMIn: Multi-Model Inference. R Package Version 1.10.0. Available online: http://cran.r-project.org/package=MuMIn (accessed on 14 January 2024).
37. Cohan, D.; Haas, S.; Roussopoulos, P.J. Decision analysis of silvicultural prescriptions and fuel management practices on an intensively managed commercial forest. *For. Sci.* **1983**, *29*, 858–870.
38. Pimont, F.; Dupuy, J.L.; Linn, R.R.; Parsons, R.; Martin-StPaul, N. Representativeness of wind measurements in fire experiments: Lessons learned from large-eddy simulations in a homogeneous forest. *Agric. For. Meteorol.* **2017**, *232*, 479–488. [CrossRef]
39. Storey, M.A.; Bedward, M.; Price, O.F.; Bradstock, R.A.; Sharples, J.J. Derivation of a Bayesian fire spread model using large-scale wildfire observations. *Environ. Model. Softw.* **2021**, *144*, 105–127. [CrossRef]
40. Kennedy, M.C.; Prichard, S.J.; McKenzie, D.; French, N.H. Quantifying how sources of uncertainty in combustible biomass propagate to prediction of wildland fire emissions. *Int. J. Wildland Fire* **2020**, *29*, 793–806. [CrossRef]
41. Skowronski, N.S.; Clark, K.L.; Duveneck, M.; Hom, J. Three-dimensional canopy fuel loading predicted using upward and downward sensing LiDAR systems. *Remote Sens. Environ.* **2011**, *115*, 703–714. [CrossRef]
42. Bastianoni, A.; Chacón, N.; Méndez, C.L.; Flores, S. Decomposition dynamics of mixed litter in a seasonally flooded forest near the Orinoco River. *Acta. Oecol.* **2015**, *64*, 21–28. [CrossRef]
43. Matthews, S.; Sullivan, A.L.; Watson, P.; Williams, R.J. Climate change, fuel, and fire behavior in a eucalypt forest. *Glob. Change Biol.* **2012**, *18*, 3212–3223. [CrossRef]
44. Liu, Y.; Liu, Y.; Fu, J.; Yang, C.E.; Dong, X.; Tian, H.; Ke, Z. Projection of future wildfire emissions in western USA under climate change: Contributions from changes in wildfire, fuel loading and fuel moisture. *Int. J. Wildland Fire* **2021**, *31*, 1–13. [CrossRef]

45. Aragoneses, E.; Chuvieco, E. Generation and mapping of fuel types for fire risk assessment. *Fire* **2021**, *4*, 59. [CrossRef]
46. Woodall, C.W.; Charney, J.; Liknes, G.; Potter, B. Linking Fuel Inventories with Atmospheric Data for Assessment of Fire Danger. In Proceedings of the Sixth Annual Forest Inventory and Analysis Symposium, Denver, CO, USA, 21–24 September 2004.

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

Wildfire Susceptibility Prediction Based on a CA-Based CCNN with Active Learning Optimization

Qiuping Yu, Yaqin Zhao *, Zixuan Yin and Zhihao Xu

College of Mechanical and Electronic Engineering, Nanjing Forestry University, Nanjing 210037, China; yuqp@njfu.edu.cn (Q.Y.); yinzixuan@njfu.edu.cn (Z.Y.); xuzhihao@njfu.edu.cn (Z.X.)
* Correspondence: zhaoyaqin@njfu.edu.cn

Abstract: Wildfires cause great losses to the ecological environment, economy, and people's safety and belongings. As a result, it is crucial to establish wildfire susceptibility models and delineate fire risk levels. It has been proven that the use of remote sensing data, such as meteorological and topographical data, can effectively predict and evaluate wildfire susceptibility. Accordingly, this paper converts meteorological and topographical data into fire-influencing factor raster maps for wildfire susceptibility prediction. The continuous convolutional neural network (CCNN for short) based on coordinate attention (CA for short) can aggregate different location information into channels of the network so as to enhance the feature expression ability; moreover, for different patches with different resolutions, the improved CCNN model does not need to change the structural parameters of the network, which improves the flexibility of the network application in different forest areas. In order to reduce the annotation of training samples, we adopt an active learning method to learn positive features by selecting high-confidence samples, which contributes to enhancing the discriminative ability of the network. We use fire probabilities output from the model to evaluate fire risk levels and generate the fire susceptibility map. Taking Chongqing Municipality in China as an example, the experimental results show that the CA-based CCNN model has a better classification performance; the accuracy reaches 91.7%, and AUC reaches 0.9487, which is 5.1% and 2.09% higher than the optimal comparative method, respectively. Furthermore, if an accuracy of about 86% is desired, our method only requires 50% of labeled samples and thus saves about 20% and 40% of the labeling efforts compared to the other two methods, respectively. Ultimately, the proposed model achieves the balance of high prediction accuracy and low annotation cost and is more helpful in classifying fire high warning zones and fire-free zones.

Keywords: wildfire susceptibility; CCNN; active learning; attention mechanism

1. Introduction

Wildfire poses a serious threat to ecosystems and human society. The frequency and intensity of wildfires are on the rise globally, which causes huge losses to the ecological environment, economy, and people's safety and belongings. To effectively manage forest areas and reduce fire risks, it has become an urgent and important task to establish efficient wildfire susceptibility models and delineate fire risk level zones.

Wildfire susceptibility is influenced by various factors, such as climate, topography, vegetation, and human factors. Among them, meteorological and topographical factors are the two most critical indicators [1–4]. Meteorological factors include precipitation, temperature, humidity, wind speed, etc., which are closely related to fire occurrence and its intensity. Topographical factors, such as elevation, slope, and aspect, play an important part in influencing the incidence, spread, and movement of wildfires [5,6]. During the past decades, meteorological and topographical information has been obtained through various channels, such as remote sensing for obtaining detailed information on vegetation

conditions, soil moisture, and fire traces [7,8], while GIS is used for mapping fire dynamics and generating fire hazard maps [9,10].

Traditional machine learning methods such as random forests [11], logistic regression [12], support vector machines [8], multilayer perceptron neural networks [13], and artificial neural networks [14,15] have been used for predicting wildfire susceptibility. However, such strategies require the selection of specific features based on a particular region, and thus, a given model can only be applicable to the task of fire susceptibility prediction in a particular scene. In addition, when dealing with complex nonlinear problems, machine learning has a relatively weak generalization capability, which hampers its ability to capture complex relationships and distributions of data.

The convolutional neural network (CNN) has found extensive applications in object recognition, target detection, robotics, and natural language processing, thanks to its powerful feature learning capabilities [16,17] in recent years. Zhang et al. [18] have attempted to use Yunnan Province as a study area and have predicted forest fire susceptibility using a CNN model. They verified the effectiveness of CNN in the task of modeling forest fire susceptibility. Although CNN can fully utilize neighborhood information and extract multilevel representations from input data, it still faces some challenges [19]. Since convolution kernels of CNN are discrete, structural parameters, including the sizes of convolutional kernels and convolutional steps, need to be readjusted when facing input images with different resolutions. To address the above limitations, a CCNN architecture was built by equipping a CNN with continuous convolutional kernels [20]. CCNN eliminates the down-sampling layers and task-dependent structural parameter settings required by current CNN architectures and is therefore suitable for tasks on data of arbitrary resolutions, dimensions, and lengths.

The attention mechanism enhances the model's ability to express features of the input images by focusing more on key information in the image. In this work, we embed the CA module into the feature extraction network of CCNN so as to enhance the model's feature expression capability. In addition, we introduce an active learning method based on two types of complementary training strategies, which can achieve high prediction performance with as few training samples as possible, thus reducing the annotation cost. The proposed model predicts wildfire susceptibility. In the method, meteorological and topographical data are processed by ArcGIS software, including rasterization, projection, resampling, and other operations, and these data are converted into raster maps of fire-influencing factors. The raster maps are then partitioned into small patches as inputs of the proposed model. As far as we know, the most similar to this work is the literature [18]; however, the aspects of our method that differ from the literature [18] are mainly in the following two points: (1) Instead of CNN structure in the literature [18], the proposed method adopts CCNN and improve feature extraction network of CCNN based on CA module. The model can process input patches of different sizes without modifying the structural parameters of the network, such as sizes of convolution kernels and convolution steps, which makes the network application more flexible. (2) We introduce an active learning method during the training process, which improves the discriminative ability of the network and reduces the annotation cost. The following points summarize the contributions of this work:

- In order to capture the spatial distribution characteristics of pixels and the correlation between local regions in the fire-influencing factor raster map, we embed the CA module into the feature extraction network of CCNN in order to aggregate different location information into channels and thereby model the global information of the fire-influencing factor raster map;
- We adopt an active learning approach based on two types of complementary learning strategies in order to improve the network's discriminative performance by prioritizing the selection of fire-influencing factor patches with high confidence. Compared with traditional learning methods, the model can achieve a balance of high prediction accuracy and low annotation cost;

- We adopt the CA-based CCNN to output the fire probability of fire-influencing factor raster patches. Instead of fixed input patches used in CNN, the proposed method allows different patch sizes without changing the structural parameters of the model, which contributes to improving the application flexibility of the wildfire susceptibility prediction model in different scenes.

2. Study Area and Data Sources

2.1. Study Area

As shown in Figure 1, Chongqing Municipality is located in southwestern China, and its geographical location is between N Lat. 28°10′ to 32°13′ and E Long. 105°11′ to 110°11′. Chongqing is densely vegetated, with more than 55 percent of the area covered by forests. Vegetation types are diverse, with subtropical evergreen broad-leaved forests and deciduous broad-leaved forests at low and middle altitudes and alpine mixed coniferous and broad-leaved forests, alpine meadows, and mixed coniferous and broad-leaved forests at high altitudes. Due to the rich vegetation, wildfires are relatively frequent. The terrain is complex, with mountains accounting for more than 70% of the topographical areas. Chongqing is well known as the 'stove' in China. In summer, the highest temperature can reach 42 °C, and the average temperature is above 32 °C; therefore, wildfires occur more frequently.

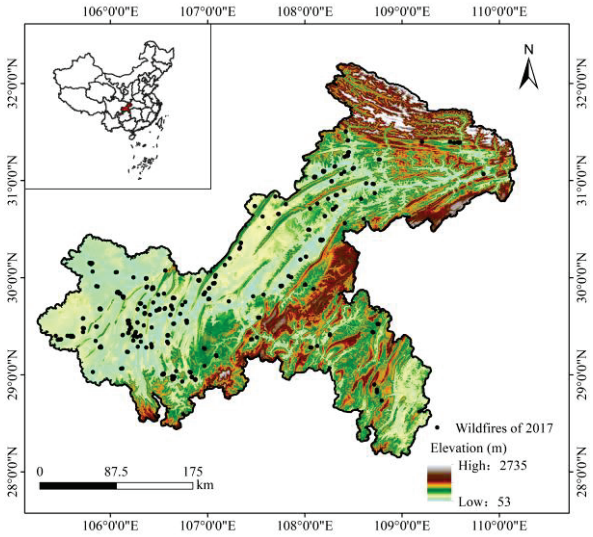


Figure 1. Location of Chongqing Municipality, in China, and distribution of wildfires in 2017.

2.2. Data Sources

Historical wildfire data sources from the Fire Information for Resource Management System website (https://firms.modaps.eosdis.nasa.gov/map/) (accessed on 6 October 2023) provided by the National Aeronautics and Space Administration. Records were derived from the fire products provided by Moderate-resolution Imaging Spectroradiometer (MODIS) from the Terra and Aqua satellites. The product covers information such as the time of fire occurrence, geographic location, and confidence of the fire point. And the spatial resolution is 1 km. We selected fire point records of years from 2012 to 2017 with frequent wildfires.

In this work, topographical and meteorological factors are used for fire susceptibility prediction. The topographical data, including elevation, slope, and aspect, are obtained from the Geospatial Data Cloud site (https://www.gscloud.cn/) (accessed on 6 October 2023) of the Computer Network Information Center, Chinese Academy of Sciences, and further calculated to obtain slope, elevation, and aspect by using the geographic data processing software. Four meteorological factors, including temperature, precipitation, wind speed, and relative humidity, are from the National Earth System Science Data Center,

National Science & Technology Infrastructure of China (http://www.geodata.cn) (accessed on 6 October 2023). The spatial resolution is 1 km, and the temporal resolution is month-by-month. As the occurrence of wildfires in Chongqing shows obvious seasonality, therefore, we apply the data in summer for modeling, namely, from June to August.

3. Methods

Figure 2 illustrates the proposed model for wildfire susceptibility prediction. The raster maps of fire-influencing factors are generated by using the software ArcGIS, and then they are integrated into one raster image, which is segmented into small image patches. The proposed model takes the patches as the inputs and predicts the fire risk probability of each patch. The CA module is embedded into the backbone network in order to capture long-range dependencies while retaining precise location information. This design can more accurately capture the spatial distribution characteristics in image patches of fire risk raster maps, thereby more comprehensively capturing the dependency relationships between features. To solve the problems of few-sample prediction and massive sample annotation, an active learning optimization approach, i.e., two types of complementary training strategy, is used to prioritize the screening of uncertain patches and patches with high confidence, where high-confidence patches are automatically assigned with pseudo-labels, which improves the model's learning capability while reducing the labeling cost.

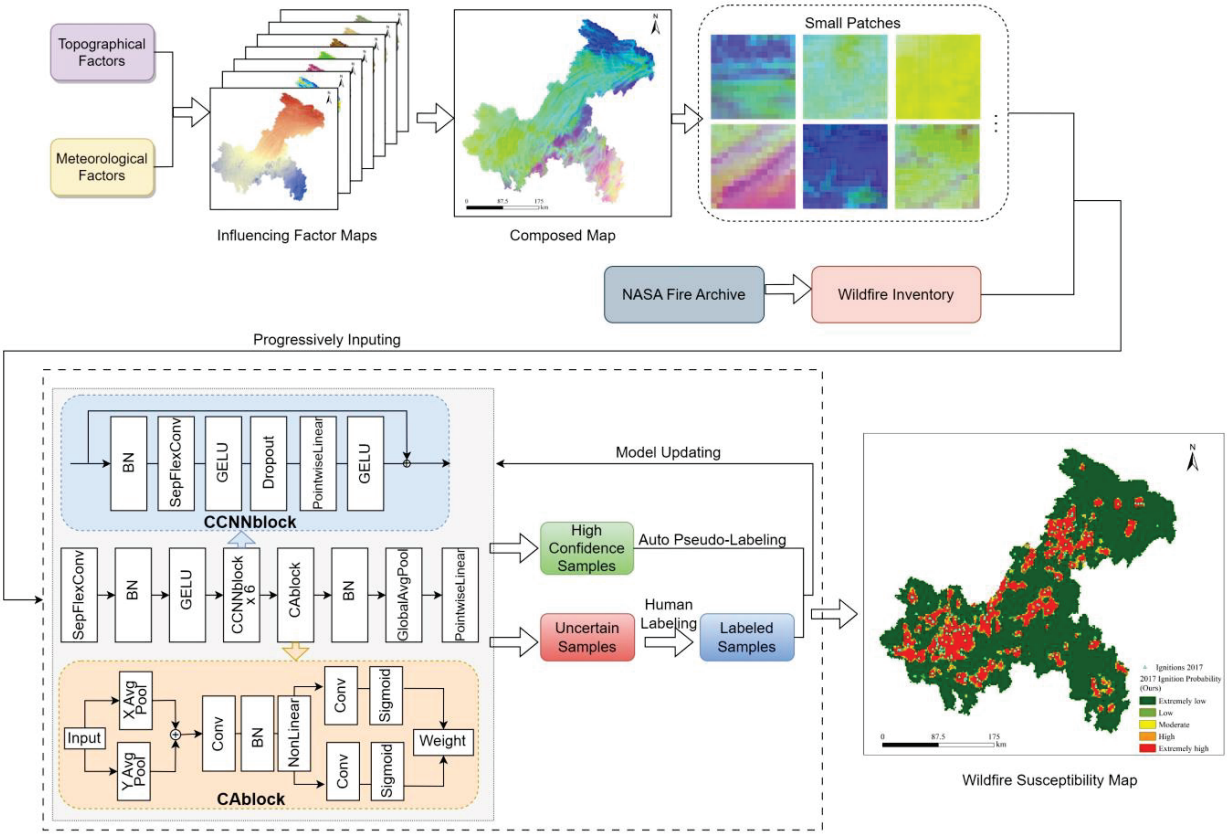


Figure 2. The model framework for the wildfire susceptibility prediction.

3.1. Generation of Fire-Influencing Factor Raster Maps

The data of seven influencing factors are utilized for wildfire susceptibility modeling. For each meteorological factor, the 3-month average value (June to August) is used to generate its raster map. We convert all influencing factor data into a raster format with the same data type, coordinate system, and resolution. Subsequently, the 7 factor raster maps of a year after the normalization operation were integrated into one raster map using the

composite bands tool. All fire-influencing factor data from 2012 to 2017 were processed in the same way to generate 6 integrated raster maps. Taking the average temperature as an example, Figure 3 shows a raster map of this influencing factor in 2017.

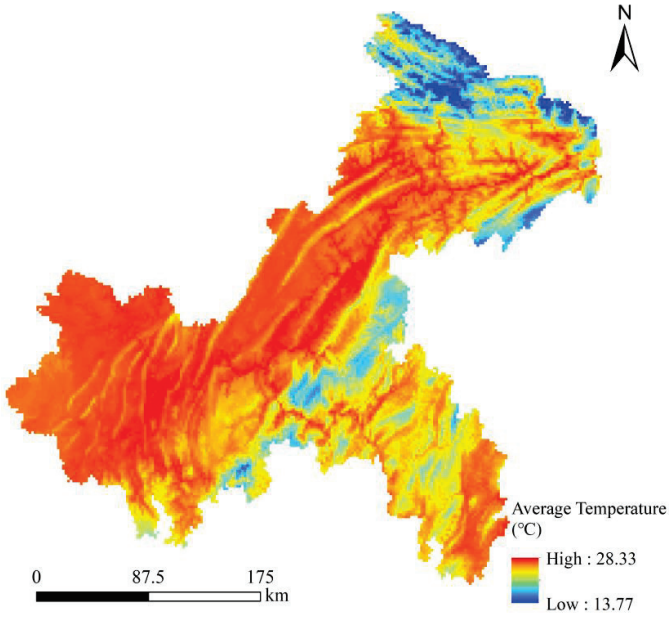


Figure 3. Raster map of the average temperature in Chongqing, China, in 2017.

3.2. The Model Input

The pixels in the raster map correspond to the geographical coordinates on the map, so the fire susceptibility prediction model evaluates the fire risk level of location points in different geographical coordinates. The fire-influencing factor data of a pixel in the raster map are approximately equal to that of its neighboring pixels; that is, the pixels in an image block usually have the same or similar fire risk level. Therefore, we segment the raster map fusing 7 fire-influencing factor information into small patches. Each patch was expressed as a 3D array with the dimension of $n \times n \times c$, where n is the patch size and c denotes the number of influencing factors.

3.3. The Structure of the CA-Based CCNN Model

In the frame of the CA-based CCNN model, the input data have 7 channels corresponding to 7 factors. After the convolution of 7 channel directions is calculated separately, the point-by-point convolution is performed. Therefore, 7 feature maps are weighted and combined in the depth direction to generate the new feature map. The convolution kernel uses SepFlexConv [20] with a stride of 1 and the same padding mode. SepFlexConv has fewer parameters and computation costs compared with the conventional convolution. Each convolutional layer is succeeded by a batch normalization (BN) layer and a Gaussian error linear unit (GELU). A total of 6 CCNN blocks are subsequently connected. The CA module is embedded after the CCNN blocks and followed by a BN layer, a global average pooling layer, and a point-wise linear connection layer.

The distinctiveness of CCNN is the introduction of a continuous convolutional kernel to handle classification tasks with different resolutions, dimensions, and lengths. The continuous convolutional kernel regards the convolution kernel as a continuous function parameterized with the help of a small neural network G_{Kernal} acting as a kernel generator network [21]. In this study, the kernel generator network is parameterized as a 3-layer MAGNet with 64 hidden units [22]. The neural network maps the coordinate $c_i \in \mathbb{R}^D$ (D is the data dimension) to the value of the convolution kernel at the corresponding position. A convolution kernel of equal size is constructed by transmitting a vector containing k coordinates $[c_i]_{i\in[1,\ldots,k]}$ through G_{Kernal}. Thereafter, the input signal is

convolved with the generated convolution kernel to obtain the output feature representation. Parameterizing the convolution kernel by a neural network can be viewed as a way of embedding data information in the neural network weights to construct continuous data representations [23,24]. The parameterization addresses the discrete characteristic of standard convolutional kernels, enabling the construction of convolutional kernels that are independent of resolution.

The CA module [25] is designed as follows: firstly, we adopt a mean pooling operation both horizontally and vertically to obtain two attention maps embedded with direction-specific information and then splice them into one attention map. Next, the spliced attention map is passed to a learning module with a convolutional layer, a BN layer, and a nonlinear connection layer. The learning module generates horizontal and vertical feature maps, which undergo a convolution operation, respectively. Ultimately, the Sigmoid activation function is employed to compute the attention weights, and the final feature map is generated by multiplying the weights into the input features.

3.4. Active Learning Training Strategy

To solve the problem of large sample annotation and few-sample predictions, we introduce an active learning method based on two types of complementary learning strategies [26]. Let $\boldsymbol{D^U}$ and $\boldsymbol{D^L}$ denote the unlabeled sample sets and labeled sample sets, respectively. At the beginning, $\boldsymbol{D^U}$ contains all the unlabeled training samples, while $\boldsymbol{D^L}$ is empty. Subsequently, two types of samples are used for training the network, i.e., uncertain samples and high-confidence samples, which are selected from $\boldsymbol{D^U}$ according to their fire probability calculated by the model.

We determine the uncertain samples based on the criterion of least confidence. The maximum value between the fire probability and non-fire probability of a sample is defined as the confidence of the sample. The confidence lc_i is calculated as

$$lc_i = \max_j p(y_i = j|\, x_i;\, w) \tag{1}$$

where x_i denotes the ith sample and j represents the jth classifier. $y_i = j$ represents the label of x_i. w are the model parameters. $p(y_i = j|\, x_i;\, w)$ denotes the probability that x_i belongs to the jth classifier. The K samples with the lowest confidence values are chosen as uncertain samples.

The high-confidence samples facilitate the learning of fire features as a way to improve the discriminative performance of the model. The model can easily discern samples with high confidence; thus, they can be automatically assigned pseudo-labels by the model. We select samples with high confidence from $\boldsymbol{D^U}$ based on the entropy. The entropy en_i is calculated as

$$en_i = -\sum_{j=1}^{m} p(y_i = j|\, x_i;\, w) \log p(y_i = j|\, x_i;\, w) \tag{2}$$

The sample x_i in $\boldsymbol{D^U}$ is chosen as high-confidence samples if the entropy en_i is less than the threshold δ. The samples with high confidence are automatically assigned pseudo-labels of 1 or 0. The initial threshold δ_0 is set to a large value to ensure high reliability of pseudo-label assignment. After each iteration, samples with high confidence are returned to $\boldsymbol{D^U}$, and their pseudo-labels are removed. After each iteration, we update the threshold.

The selected uncertain samples need to be manually labeled and then added to the labeled sample set $\boldsymbol{D^L}$, while the model automatically assigns pseudo-labels to samples with high confidence. As the number of iterations increases, more uncertain samples are selected; at the same time, the model has higher performance, which made the model achieve a balance between the number of sample annotations and model performance. Since we automatically assign pseudo-labels to high-confidence samples without manual labeling, the cost of manual labeling is greatly reduced.

4. Experiments

In this section, the configurations of experiments and parameters of the training process are provided, and then the metrics used in this work are introduced.

4.1. Implement Details

Our framework is implemented using PyTorch Lightning 1.6.4. The configurations of the computer are as follows: a GPU (RTX 2080Ti) and a CPU (Intel i9-9900K). We employ AdamW [27] as the optimizer with a weight decay set to 0. The default learning rate is 0.01. In addition, we introduce a cosine annealing learning rate scheduler [28] along with a linear learning rate warm-up stage of 10 epochs for optimization. The dropout rate is set to 0.15 to prevent overfitting.

The fire points listed on the inventory from NASA represent the origin or center of a fire region rather than the actual fire region; therefore, we adopted buffer analysis [29] to expand fire point data into a fire region. Drawing on the literature [18], a buffer zone was set within 5 km around the fire point, and pixels inside the buffer zone were labeled as 1 (i.e., fire), while pixels outside the buffer zone were labeled as 0 (i.e., non-fire). All fire point data from 2012 to 2017 were processed to generate six ignition raster maps. The annual raster maps from 2012 to 2016 are used for model training and are divided into $n \times n \times 7$ 3D patches, where n is set to 15, 25, or 35. A total of 5000 patches, including 2500 fire samples and 2500 non-fire samples, are randomly chosen for model training. The training and validation sets consist of 4000 and 1000 samples, respectively. The data from 2017 served as the testing set. A total of 10% of the labeled samples in the training set are randomly selected for network initialization.

4.2. Metrics

In this work, the prediction for wildfire susceptibility is realized by a binary classification method. Thus, we use metrics such as sensitivity, specificity, positive predictive value (PPV for short), and negative predictive value (NPV for short) and accuracy to evaluate the model performance [9]. The calculation of the five metrics is as follows:

$$\begin{aligned} &\text{Sensitivity} = \frac{TP}{TP+FN} \\ &\text{Specificity} = \frac{TN}{FP+TN} \\ &\text{PPV} = \frac{TP}{FP+TP} \\ &\text{NPV} = \frac{TN}{FN+TN} \\ &\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \end{aligned} \tag{3}$$

The overall performance of the model is assessed using the area under the curve of the receiver operating characteristic (i.e., ROC-AUC) [30–32]. A higher AUC value, closer to 1, indicates better prediction performance [33]. AUC is calculated as follows:

$$AUC = \frac{1 + \text{TPR} - \text{FPR}}{2} \tag{4}$$

5. Results

In this section, we verify that our method achieves high prediction accuracy at different patch sizes. At the same time, our method is compared with several state-of-the-art methods, and wildfire susceptibility maps are generated, respectively. Finally, the effectiveness of the two types of complementary learning strategies in reducing training samples is verified.

5.1. Ablation Study

We conducted experiments with different sizes of patches, and as shown in Table 1, the best prediction performance was achieved at the patch size of 25 $\times$ 25; hence, the 25 $\times$ 25 patch was chosen for the subsequent experiments in this work.

Table 1. Performance comparison with different input patch sizes.

Input Size		**15 × 15**	**25 × 25**	**35 × 35**
Training	TP	2005	2005	2000
	TN	1992	1993	1988
	FP	2	1	6
	FN	1	1	6
	Sensitivity (%)	99.95	99.95	99.70
	Specificity (%)	99.90	99.95	99.70
	PPV (%)	99.90	99.95	99.70
	NPV (%)	99.95	99.95	99.70
	Accuracy (%)	99.93	99.95	99.70
Validating	TP	466	468	472
	TN	447	449	440
	FP	59	57	66
	FN	28	26	22
	Sensitivity (%)	94.33	94.74	95.55
	Specificity (%)	88.34	88.74	86.96
	PPV (%)	88.76	89.14	87.73
	NPV (%)	94.11	94.53	95.24
	Accuracy (%)	91.30	91.70	91.20

CNN is more suitable for classifying such small-sized input patches than other deep learning models; thus, we compared our method with the CNN-based method proposed by [18]. Additionally, we evaluated our method against CCNN, CCNN + SE, and CCNN + CBAM models. As shown in Table 2, our method outperforms these models across several metrics, including sensitivity, specificity, PPV, NPV, and accuracy. Figure 4 demonstrates the ROC curves for the five methods on the testing set. The AUC of our method is 0.9487, which is 2.09% higher than the CNN-based method and 0.73% higher than the CCNN method. Furthermore, the AUC of our method exceeds that of the CCNN model using two other additional attentional mechanisms. This indicates a better overall fit between the CA-based CCNN wildfire susceptibility prediction model and the testing set.

Table 2. Performance comparison of our method with state-of-the-art methods.

Models		**Ours**	**CNN-Based Method [18]**	**CCNN**	**CCNN + SE**	**CCNN + CBAM**
Training	TP	2005	1992	2001	2000	1997
	TN	1993	1918	1978	1984	1983
	FP	1	71	16	10	11
	FN	1	19	5	6	9
	Sensitivity (%)	99.95	99.06	99.75	99.70	99.55
	Specificity (%)	99.95	96.43	99.20	99.50	99.45
	PPV (%)	99.95	96.56	99.21	99.50	99.45
	NPV (%)	99.95	99.02	99.75	99.70	99.55
	Accuracy (%)	99.95	97.75	99.48	99.60	99.50
Validating	TP	468	437	464	463	460
	TN	449	429	436	448	444
	FP	57	82	70	58	62
	FN	26	52	30	31	34
	Sensitivity (%)	94.74	89.37	93.93	93.72	93.12
	Specificity (%)	88.74	83.95	86.17	88.54	87.75
	PPV (%)	89.14	84.2	86.89	88.87	88.12
	NPV (%)	94.53	89.19	93.56	93.53	92.89
	Accuracy (%)	91.70	86.60	90.00	91.10	90.40

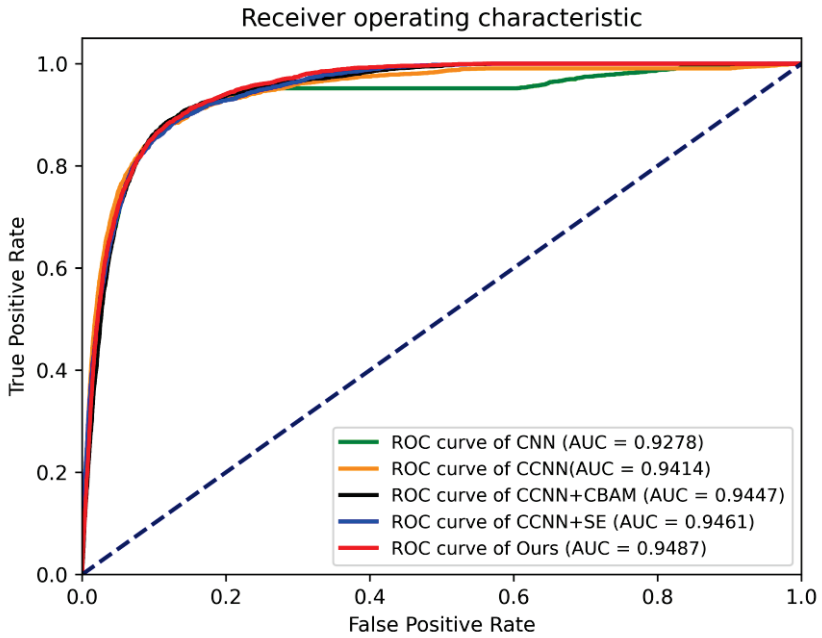


Figure 4. ROC curves and AUC of the five methods.

5.2. Comparison with State-of-the-Art Machine Learning Methods

In addition to CNN, we also use machine learning methods for comparative experiments. The experimental results are shown in Table 3 and Figure 5, where our method shows higher sensitivity (94.74%), specificity (88.74%), PPV (89.14%), and NPV (94.53%) than the four machine learning methods on the validation set. And the overall accuracy of the method in this work reaches 91.7%, which is more than 15% higher than the machine learning methods. Similarly, CNN-based methods have achieved high performance. These results indicate that the raster-based scheme may outperform traditional machine learning methods in wildfire susceptibility tasks.

Table 3. Performance comparison of our method with machine learning methods.

Models		Ours	CNN	RF	Decision Tree	MLP	SVM
Training	TP	2005	1992	2003	1695	1530	1611
	TN	1993	1918	1985	1742	1555	1558
	FP	1	71	4	247	434	431
	FN	1	19	8	316	481	400
	Sensitivity (%)	99.95	99.06	99.60	84.29	76.08	80.11
	Specificity (%)	99.95	96.43	99.80	87.58	78.18	78.33
	PPV (%)	99.95	96.56	99.80	87.28	77.90	78.89
	NPV (%)	99.95	99.02	99.60	84.65	76.38	79.57
	Accuracy (%)	99.95	97.75	99.70	85.93	77.13	79.23
Validating	TP	468	437	353	331	346	358
	TN	449	429	379	382	396	382
	FP	57	82	132	129	115	129
	FN	26	52	136	158	143	131
	Sensitivity (%)	94.74	89.37	72.19	67.69	70.76	73.21
	Specificity (%)	88.74	83.95	74.17	74.76	77.50	74.76
	PPV (%)	89.14	84.2	72.78	71.96	75.05	73.51
	NPV (%)	94.53	89.19	73.59	70.74	73.47	74.46
	Accuracy (%)	91.70	86.60	73.20	71.30	74.20	74.00

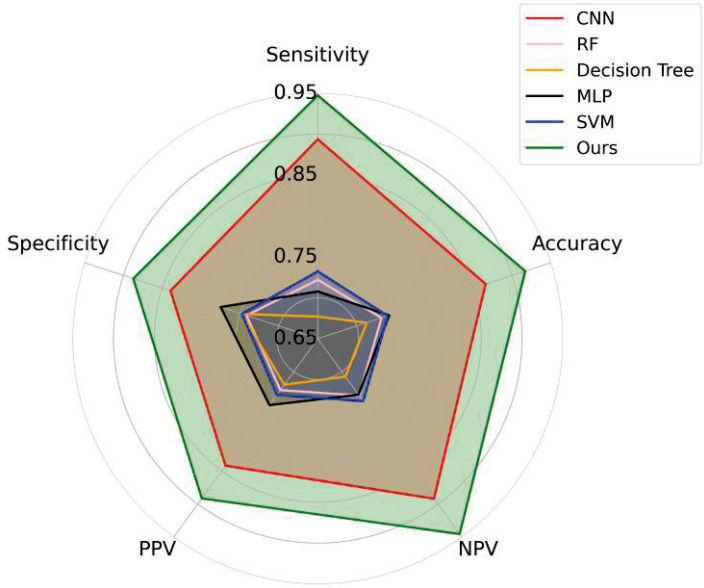


Figure 5. Radar maps of metrics derived from six different models on the validation set.

5.3. Wildfire Susceptibility Maps

Wildfire susceptibility was predicted using the testing set. We utilize fire probability predicted by the model to generate the wildfire susceptibility map and then classify wildfire susceptibility into five classes: extremely low, low, moderate, high, and extremely high, by selecting the Natural Breaks method in ArcGIS for the threshold setting.

Figure 6 displays wildfire susceptibility maps generated by our method and five state-of-the-art methods, respectively. In these maps, blue triangular arrows represent fire points. The extremely high fire susceptibility approximately indicates the occurrence of a fire. The areas with extremely high fire susceptibility predicted by our method are consistent with the actual fire location. However, the MLP, SVM, and RF methods incorrectly classify these areas as low- and medium-susceptibility zones. In comparison with the CNN-based method, the susceptibility map predicted by our method offers slightly more accurate delineation in the extremely low and extremely high susceptibility classes.

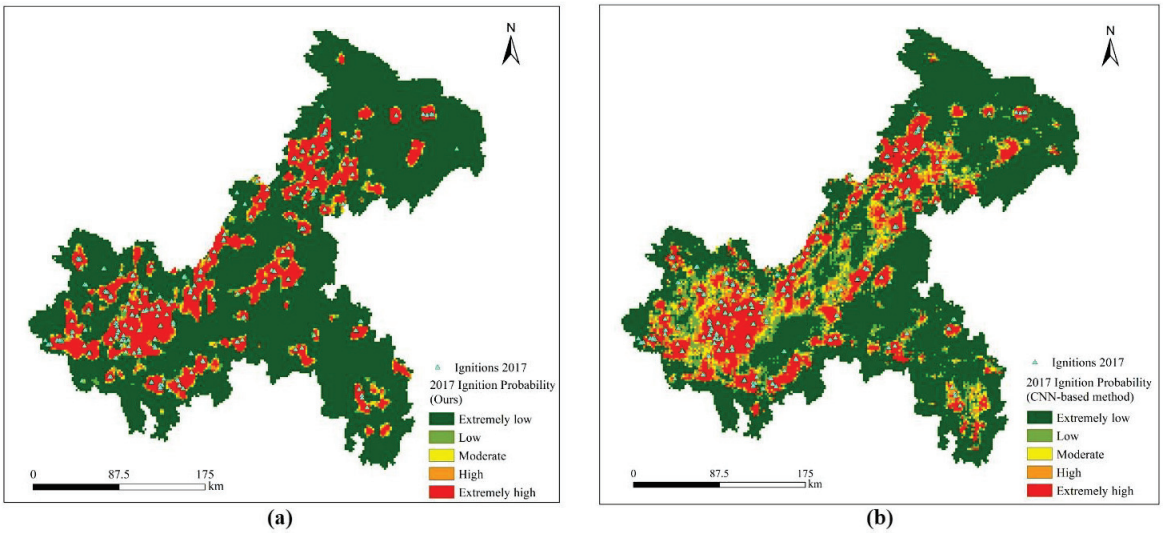


Figure 6. *Cont.*

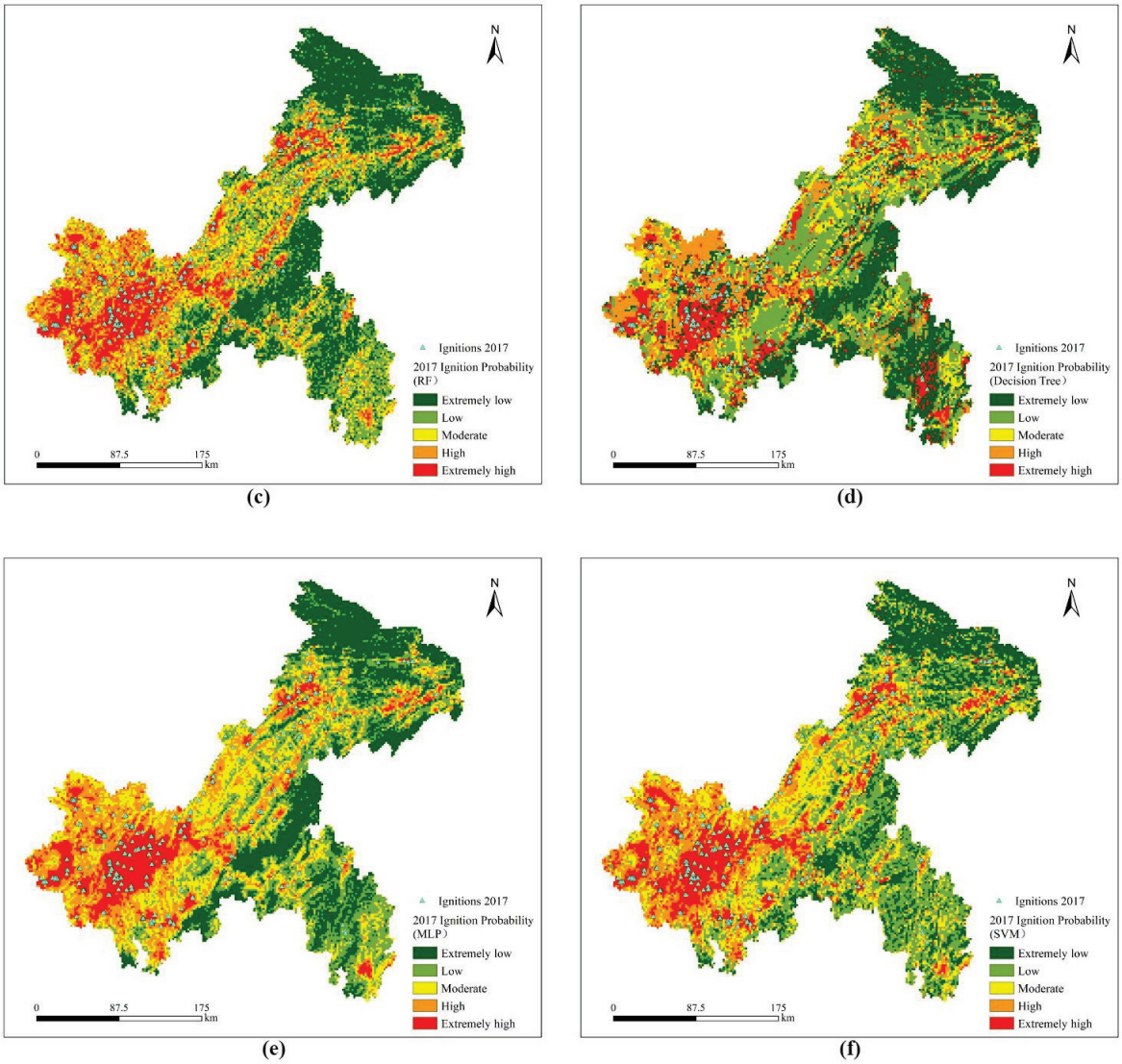


Figure 6. Wildfire susceptibility maps derived from different methods for Chongqing Municipality, China, in 2017. (**a**) Our method; (**b**) CNN-based method; (**c**) RF; (**d**) Decision Tree; (**e**) MLP; (**f**) SVM.

5.4. Verification of Training Efficiency of Two Types of Complementary Learning Strategy

In order to demonstrate the training efficiency of the active learning strategy used in this work, two training methods, i.e., randomly selecting samples and only selecting uncertain samples [34], were used for the comparison. The parameters of the active learning strategy are set as follows: $\delta_0 = 0.05$ and $K = 400$.

Figure 7 displays the classification accuracy curve of these three methods under different percentages of labeled samples. The proposed method exhibits superior prediction accuracy compared to the other two methods with the same percentage of labeled samples, particularly when the percentage of labeled samples is low (20–60%), with the superiority becoming more pronounced.

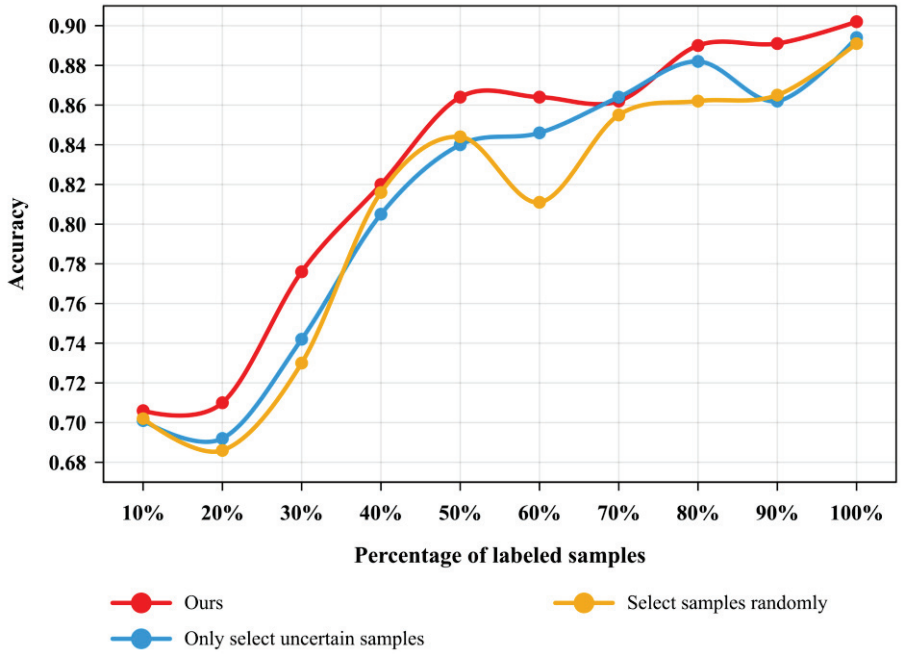


Figure 7. Classification accuracy with different percentages of labeled samples.

6. Discussions

6.1. Fire-Influencing Factors and Fire Susceptibility Prediction

Overall, wildfires in Chongqing occur mainly in forested areas. In particular, the southwest region of Chongqing has low elevation and high temperature in summer, which heightens the likelihood of fire occurrence. Wind speed increases with the increment of slope gradient in the forest areas in the north–central region. Therefore, these two regions are prone to frequent fires, which is consistent with the prediction results of our method. Therefore, there is a direct causal relationship between fire susceptibility and the selected fire-influencing factors, including topographical and meteorological factors.

6.2. Advantages of CCNN

Compared with CNN, CCNN applies a continuous convolutional kernel, enabling the creation of convolutional kernels independent of resolution and thus increasing the flexibility of the model to be applied in different regions. As seen in Table 2 and Figure 6, the model parameters, such as the size of the convolutional kernel and convolutional strides, were determined after training. The superior performance of the proposed method is not affected by the size of input patches due to the introduction of continuous convolutional kernels into CCNN.

6.3. The Basis for CA Module Selection

Currently, the widely employed attention mechanisms include the Squeeze-and-Excitation (SE) mechanism, CBAM mechanism, and CA mechanism. We introduce the CA mechanism in order to capture the spatial distribution characteristics of pixels in the fire-influencing factor raster map. As shown in Table 2, our method has higher prediction accuracy than CCNN + CBAM and CCNN + SE. Because CA aggregates different location information into channels of the network, the correlation between local regions is captured [25]. Due to discrete pixels in one patch, we utilize CA to model the long-range dependencies between pixels. SE mechanism focuses solely on the encoded inter-channel information, overlooking the significance of positional information [35]. CBAM is limited to capturing local relationships [36]. Therefore, CA is selected to model the global information in the patches of fire-influencing factor raster map.

6.4. Labeling Cost

By analyzing Figure 7, we find that if an accuracy of about 86% is desired, the methods of only selecting uncertain samples and randomly selecting samples require about 70% and 90% labeled training samples, respectively. In contrast, our method requires only 50% labeled samples, saving about 20% and 40% labeling efforts. Thus, the proposed method can save the labeling cost under the demand for relatively high prediction accuracy. We also observe that the method of only selecting uncertain samples has slightly higher prediction accuracy than the other two comparative methods when the proportion of labeled training samples is 70%. This is because the best training samples are accidentally chosen; that is, this excellent performance is random and uncertain.

6.5. Wildfire Susceptibility Map

As shown in Figure 6, we mark the fire risk level on the map of Chongqing. Therefore, the fire susceptibility map includes geographic information of the predicted area, achieving the visualization of fire risk prediction. In addition, as seen in Figure 2, after the fire-influencing factor raster map is input into the model, it is segmented into small patches so that we can simultaneously output the prediction results for all regions of Chongqing. On the contrary, traditional machine learning methods can only predict the fire risk of a small area and output the fire risk value of that area.

7. Conclusions

In this work, we proposed a CA-based CCNN model with active learning optimization for wildfire susceptibility prediction. Due to the CCNN architecture, the model is compatible with different input patch sizes, which contributes to applying the pre-trained fire susceptibility prediction model to different forest areas. In addition, the model adopted CA to build the long-range dependencies for discrete pixels in small fire-influencing factor patches. Experimental results indicated that the CCNN + CA model had higher prediction accuracy than the comparative state-of-the-art methods. Furthermore, compared with traditional machine learning methods, the model saved the labeling cost by using an active learning method. It was worth mentioning that we utilized the model to generate a fire susceptibility map that visualized the fire risk level with geographic information. In future work, we will take into account the influence of anthropogenic activities on wildfire occurrence to reveal the interests and conflicts between wildfires and human activities. We will also draw lessons from the continuous convolution approach of CCNN into lightweight convolutional networks so as to further reduce complexity.

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

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Datasets are available from the corresponding author upon reasonable request.

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

References

1. Bui, D.T.; Le, H.V.; Hoang, N.D. GIS-based spatial prediction of tropical forest fire danger using a new hybrid machine learning method. *Ecol. Inform.* **2018**, *48*, 104–116. [CrossRef]
2. Sachdeva, S.; Bhatia, T.; Verma, A.K. GIS-based evolutionary optimized Gradient Boosted Decision Trees for forest fire susceptibility mapping. *Nat. Hazards* **2018**, *92*, 1399–1418. [CrossRef]

3. Razavi-Termeh, S.V.; Sadeghi-Niaraki, A.; Choi, S.M. Ubiquitous GIS-Based Forest Fire Susceptibility Mapping Using Artificial Intelligence Methods. *Remote Sens.* **2020**, *12*, 1689. [CrossRef]
4. McEvoy, A.; Nielsen-Pincus, M.; Holz, A.; Catalano, A.J.; Gleason, K.E. Projected Impact of Mid-21st Century Climate Change on Wildfire Hazard in a Major Urban Watershed outside Portland, Oregon USA. *Fire* **2020**, *3*, 70. [CrossRef]
5. Nguyen, N.T.; Dang, B.T.N.; Pham, X.C.; Nguyen, H.T.; Bui, H.T.; Hoang, N.D.; Bui, D.T. Spatial pattern assessment of tropical forest fire danger at Thuan Chau area (Vietnam) using GIS-based advanced machine learning algorithms: A comparative study. *Ecol. Inform.* **2018**, *46*, 74–85. [CrossRef]
6. Cicione, A.; Gibson, L.; Wade, C.; Spearpoint, M.; Walls, R.; Rush, D. Towards the Development of a Probabilistic Approach to Informal Settlement Fire Spread Using Ignition Modelling and Spatial Metrics. *Fire* **2020**, *3*, 67. [CrossRef]
7. Maeda, E.E.; Formaggio, A.R.; Shimabukuro, Y.E.; Arcoverde, G.F.B.; Hansen, M.C. Predicting forest fire in the Brazilian Amazon using MODIS imagery and artificial neural networks. *Int. J. Appl. Earth Obs. Geoinf.* **2009**, *11*, 265–272. [CrossRef]
8. Mohajane, M.; Costache, R.; Karimi, F.; Pham, Q.B.; Essahlaoui, A.; Nguyen, H.; Laneve, G.; Oudija, F. Application of remote sensing and machine learning algorithms for forest fire mapping in a Mediterranean area. *Ecol. Indic.* **2021**, *129*, 107869. [CrossRef]
9. Bui, D.T.; Bui, Q.T.; Nguyen, Q.P.; Pradhan, B.; Nampak, H.; Trinh, P.T. A hybrid artificial intelligence approach using GIS-based neural-fuzzy inference system and particle swarm optimization for forest fire susceptibility modeling at a tropical area. *Agr. Forest Meteorol.* **2017**, *233*, 32–44. [CrossRef]
10. Sivrikaya, F.; Küçük, O. Modeling forest fire risk based on GIS-based analytical hierarchy process and statistical analysis in Mediterranean region. *Ecol. Inform.* **2022**, *68*, 15. [CrossRef]
11. Pourtaghi, Z.S.; Pourghasemi, H.R.; Aretano, R.; Semeraro, T. Investigation of general indicators influencing on forest fire and its susceptibility modeling using different data mining techniques. *Ecol. Indic.* **2016**, *64*, 72–84. [CrossRef]
12. 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. [CrossRef]
13. Vasconcelos, M.J.P.D.; Silva, S.; Tome, M.; Alvim, M.; Pereira, J.M.C. Spatial prediction of fire ignition probabilities: Comparing logistic regression and neural networks. *Photogramm. Eng. Remote Sens.* **2001**, *67*, 73–81.
14. Dimuccio, L.A.; Ferreira, R.; Cunha, L.; de Almeida, A.C. Regional forest-fire susceptibility analysis in central Portugal using a probabilistic ratings procedure and artificial neural network weights assignment. *Int. J. Wildland Fire* **2011**, *20*, 776–791. [CrossRef]
15. Satir, O.; Berberoglu, S.; Donmez, C. Mapping regional forest fire probability using artificial neural network model in a Mediterranean forest ecosystem. *Geomat. Nat. Hazards Risk* **2016**, *7*, 1645–1658. [CrossRef]
16. Dong, Z.; Pei, M.T.; He, Y.; Liu, T.; Dong, Y.M.; Jia, Y.D. Vehicle Type Classification Using Unsupervised Convolutional Neural Network. In Proceedings of the 22nd International Conference on Pattern Recognition (ICPR), Swedish Soc Automated Image Anal, Stockholm, Sweden, 24–28 August 2014; pp. 172–177.
17. Yan, X.; Liu, Y.; Xu, Y.; Jia, M. Multistep forecasting for diurnal wind speed based on hybrid deep learning model with improved singular spectrum decomposition. *Energy Convers. Manag.* **2020**, *225*, 113456. [CrossRef]
18. Zhang, G.L.; Wang, M.; Liu, K. Forest Fire Susceptibility Modeling Using a Convolutional Neural Network for Yunnan Province of China. *Int. J. Disaster Risk Sci.* **2019**, *10*, 386–403. [CrossRef]
19. Huang, Z.; Zhu, T.; Li, Z.; Ni, C. Non-Destructive Testing of Moisture and Nitrogen Content in Pinus Massoniana Seedling Leaves with NIRS Based on MS-SC-CNN. *Appl. Sci.* **2021**, *11*, 2754. [CrossRef]
20. Knigge, D.M.; Romero, D.W.; Gu, A.; Gavves, E.; Bekkers, E.; Tomczak, J.M.; Hoogendoorn, M.; Sonke, J. Modelling Long Range Dependencies in N-D: From Task-Specific to a General Purpose CNN. *arxiv* **2023**, arXiv:2301.10540.
21. Romero, D.W.; Kuzina, A.; Bekkers, E.J.; Tomczak, J.M.; Hoogendoorn, M. CKConv: Continuous Kernel Convolution For Sequential Data. *arxiv* **2021**, arXiv:2102.02611.
22. Romero, D.W.; Bruintjes, R.J.; Tomczak, J.M.; Bekkers, E.J.; Hoogendoorn, M.; Gemert, J.C.V. FlexConv: Continuous Kernel Convolutions with Differentiable Kernel Sizes. *arxiv* **2021**, arXiv:2110.08059.
23. Fathony, R.; Sahu, A.K.; Willmott, D.; Kolter, J.Z. Multiplicative filter networks. In Proceedings of the 9th International Conference on Learning Representations (ICLR), Virtual Event, Austria, 3–7 May 2021.
24. Park, J.J.; Florence, P.; Straub, J.; Newcombe, R.; Lovegrove, S. DeepSDF: Learning Continuous Signed Distance Functions for Shape Representation. In Proceedings of the 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 165–174.
25. Hou, Q.B.; Zhou, D.Q.; Feng, J.S. Coordinate Attention for Efficient Mobile Network Design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Electr Network, Virtual, 19–25 June 2021; pp. 13708–13717.
26. Wang, K.Z.; Zhang, D.Y.; Li, Y.; Zhang, R.M.; Lin, L. Cost-Effective Active Learning for Deep Image Classification. *IEEE Trans. Circuits Syst. Video Technol.* **2017**, *27*, 2591–2600. [CrossRef]
27. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. *arxiv* **2017**, arXiv:1711.05101.
28. Loshchilov, I.; Hutter, F. SGDR: Stochastic Gradient Descent with Warm Restarts. *arxiv* **2016**, arXiv:1608.03983.
29. Cao, Y.X.; Wang, M.; Liu, K. Wildfire Susceptibility Assessment in Southern China: A Comparison of Multiple Methods. *Int. J. Disaster Risk Sci.* **2017**, *8*, 164–181. [CrossRef]
30. Fadhillah, M.F.; Lee, S.; Lee, C.W.; Park, Y.C. Application of Support Vector Regression and Metaheuristic Optimization Algorithms for Groundwater Potential Mapping in Gangneung-si, South Korea. *Remote Sens.* **2021**, *13*, 1196. [CrossRef]

31. Hakim, W.L.; Achmad, A.R.; Lee, C.W. Land Subsidence Susceptibility Mapping in Jakarta Using Functional and Meta-Ensemble Machine Learning Algorithm Based on Time-Series InSAR Data. *Remote Sens.* **2020**, *12*, 3627. [CrossRef]
32. Han, J.; Nur, A.S.; Syifa, M.; Ha, M.; Lee, C.W.; Lee, K.Y. Improvement of Earthquake Risk Awareness and Seismic Literacy of Korean Citizens through Earthquake Vulnerability Map from the 2017 Pohang Earthquake, South Korea. *Remote Sens.* **2021**, *13*, 1365. [CrossRef]
33. Hosmer, D.W.; Lemeshow, S.; Sturdivant, R.X. *Applied Logistic Regression*, 3rd ed.; Wiley: Hoboken, NJ, USA, 2013.
34. Demir, B.; Bruzzone, L. A Novel Active Learning Method in Relevance Feedback for Content-Based Remote Sensing Image Retrieval. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 2323–2334. [CrossRef]
35. Wang, H.; Zhu, Y.; Green, B.; Adam, H.; Yuille, A.; Chen, L.C. Axial-DeepLab: Stand-Alone Axial-Attention for Panoptic Segmentation. *arxiv* **2020**, arXiv:2003.07853.
36. Hou, Q.B.; Zhang, L.; Cheng, M.M.; Feng, J.S. Strip Pooling: Rethinking Spatial Pooling for Scene Parsing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Electr Network, Seattle, WA, USA, 14–19 June 2020; pp. 4002–4011.

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

Fire-RPG: An Urban Fire Detection Network Providing Warnings in Advance

Xiangsheng Li and Yongquan Liang *

College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao 266590, China; lxs@sdust.edu.cn
* Correspondence: lyq@sdust.edu.cn

Abstract: Urban fires are characterized by concealed ignition points and rapid escalation, making the traditional methods of detecting early stage fire accidents inefficient. Thus, we focused on the features of early stage fire accidents, such as faint flames and thin smoke, and established a dataset. We found that these features are mostly medium-sized and small-sized objects. We proposed a model based on YOLOv8s, Fire-RPG. Firstly, we introduced an extra very small object detection layer to enhance the detection performance for early fire features. Next, we optimized the model structure with the bottleneck in GhostV2Net, which reduced the computational time and the parameters. The Wise-IoUv3 loss function was utilized to decrease the harmful effects of low-quality data in the dataset. Finally, we integrated the low-cost yet high-performance RepVGG block and the CBAM attention mechanism to enhance learning capabilities. The RepVGG block enhances the extraction ability of the backbone and neck structures, while CBAM focuses the attention of the model on specific size objects. Our experiments showed that Fire-RPG achieved an mAP of 81.3%, an improvement of 2.2%. In addition, Fire-RPG maintained high detection performance across various fire scenarios. Therefore, our model can provide timely warnings and accurate detection services.

Keywords: urban fire; YOLOv8; early fire warning; small object

1. Introduction

Fires have been a significant issue for a long time, bringing economic loss and endangering the safety of persons at the scene and the firefighters engaged in the rescue efforts. According to the China National Fire Rescue Bureau [1–5], the data on fires have historically been collected based on four indicators: the total number of fires, the number of deaths in total, the number of fatalities in residential fires, and economic losses.

According to Figure 1a, the total number of domestic fires in 2022 grew by approximately 2.5 times compared to those in 2018. In terms of economic loss, the financial loss from fires in 2022 was nearly double that of 2018. These data show that with the development of the economy, potential fire hazards have significantly increased, leading to more destructive fire incidents and more significant property damage. In Figure 1b, it is evident that over half of all fatalities were caused by residential fires, proving that these fires pose significant danger to residents' safety.

The causes of urban fire incidents [6] are relatively complex, including the improper use of fire and the disposal of flammable items. Moreover, the complex urban environment makes it challenging to detect fires in time and facilitates the rapid spread of fires. Therefore, urban fire accidents are characterized by concealed ignition points and rapid escalation.

Manual inspection is the earliest method of fire detection, involving the arrangement of personnel to inspect potential fire sites regularly. However, with the development of society, the limitations of manual inspection have become increasingly apparent, such as long response times and difficulties in the early stages of fires, making it unsuitable for current needs.

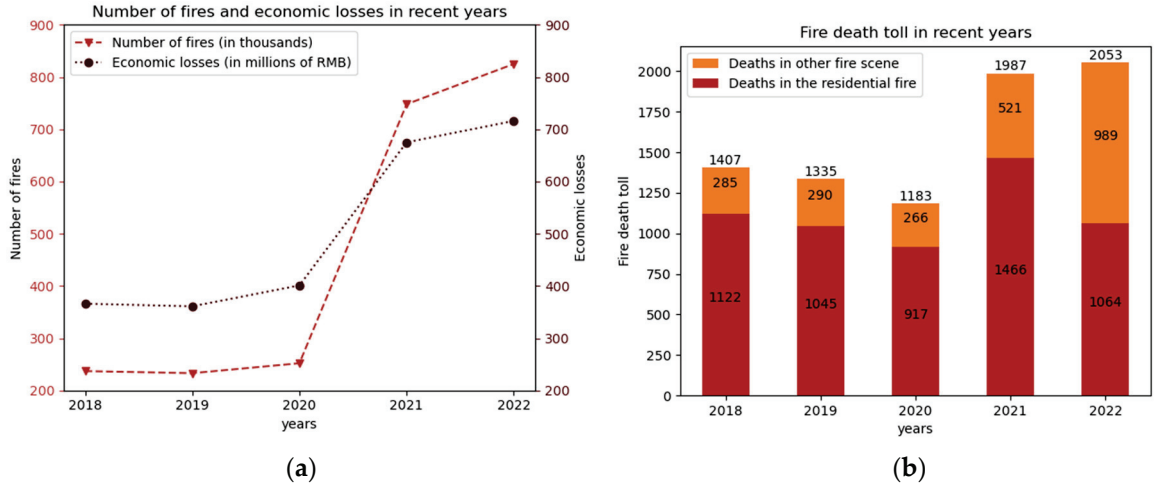


Figure 1. National fire situation in recent years. (**a**) Number of fires and economic losses; (**b**) Fire death toll.

The sensor-detection method [7] requires sensors installed in advance and detects fire accidents through smoke, flames, or high temperatures. This method has many limitations: a single sensor can only detect the presence of a fire, not the severity of the fire or the specific ignition point, and it may even produce false alarms. In addition, fire sensors have a restricted detection range, requiring proximity to the fire to activate an alarm. As a result, sensors may be damaged by the fire, leading to higher maintenance costs for the fire-detection system.

The process of urbanization has also led to the widespread deployment of CCTV (closed-circuit television) monitoring systems throughout streets and inside buildings, making it possible to monitor fire hazards at a low cost. Fire-detection methods can be built into monitoring systems [8], utilizing artificial intelligence based on computer vision technology to detect fires.

The initial intelligent fire-detection methods [9] utilized image processing technology. These methods involve extracting features from images and using machine learning algorithms to produce a model capable of detecting fires. Features like color, edges, and texture can help to identify flame objects in an image. However, due to the complexity of real-world fire scenes, the models lack the generalization ability and cannot handle natural fire scenes.

Advancements in deep learning have led to significant progress in object-detection algorithms utilizing CNNs (convolutional neural networks). Compared to traditional detection algorithms, the CNN extracts more object features than conventional methods, thus significantly improving accuracy. CNN object-detection algorithms have two categories: two-stage and one-stage. The difference between them is the detection steps. Two-stage algorithms typically begin by generating region proposals through the RPN (region proposal network) before using the CNN to identify objects in the image and assign objects to specific categories, for example, Mask-RCNN [10] and Faster-RCNN [11]. However, one-stage algorithms directly detect the category and location of objects by using CNNs, such as SSD [12], RetinaNet [13], and YOLO [14]. Hence, the inference time of one-stage algorithms is shorter than that of two-stage algorithms.

Among the one-stage algorithms, the YOLO series of algorithms stands out for their simple structure, high accuracy, and real-time detection capabilities. Li [15] found that YOLOv3 has a higher average precision than other object-detection algorithms and possesses robust detection capabilities by comparing mainstream object-detection algorithms. The experiment carried out by Li demonstrated that YOLO-series networks perform excellently in fire-detection tasks.

Fire accidents typically progress from the initial ignition to fire escalation. Therefore, fire accidents usually go through a phase that is easy to control but difficult to detect. In order to detect this phase as early as possible, researchers typically choose to reduce the parameters and the required inference time of the model while adding high-performance modules to enhance detection capabilities.

For example, Wang [16] introduced an extra very small object detection layer into the YOLOv5 network and used ghost convolutions to reduce the computational cost caused by this layer. The network achieved improved detection capabilities for small objects, such as faint flames, with a relatively low computational cost. Lin [17] introduced a coordinate attention mechanism into the network to focus the attention of the model on forest fires and improve the learning ability for key objects. Zhao [18] found that the backbone structure of YOLOv3 Darknet-53 has the disadvantage of imbalances in depth, width, and resolution. They replaced Darknet-53 with an improved EfficientNet and proposed an improved model based on YOLOv3 with faster speed, fewer parameters, and greater detection capabilities. Zhang et al. [19] proposed a ship fire-detection network based on the YOLOv8n network. The team introduced an extra object-detection layer and used spatial and channel reconstruction convolutions to reduce redundant features. These modules can improve the detection capabilities of small objects. However, the harsh hardware conditions in the ship environment constrained the detection performance of the extra detection layer. Chen [20] utilized ghost shuffle convolutions in the YOLOv7 network deployed on UAV platforms. Ghost shuffle convolution can improve the network's expression capabilities and decrease the parameters. Chen also replaced the CIoU loss function with the highly fitting SIoU loss function, accelerating the convergence speed and improving the detection capabilities of the model without additional cost.

Many researchers have proposed excellent detection algorithms. However, the characteristics of urban fire accidents, like concealed ignition points and rapid escalation, still make it difficult for these algorithms to provide timely warnings. Current fire detection networks do well in the detection of strong flames and dense smoke, but their detection performance for faint flames and thin smoke is not reliable. Faint flames and thin smoke are often considered early warning signals of fire accidents because they usually appear in the initial stages. These objects should be given significant attention, but they are usually small and difficult to detect, and the complex urban environment makes timely detection even more challenging.

Hence, detecting these objects earlier without a huge computational burden on the network is a major issue for urban fire detection systems. This study has established an urban fire dataset and proposed an urban fire detection network based on YOLOv8s [21] named Fire-RPG, where RPG stands for Reparameterization and GhostV2. This network focuses on the early warning signals of fire accidents and aims at high accuracy without much computational cost.

This study makes the following contributions:

1. RepVGG replaces complex structures with a simple equivalent structure in the inference phase, improving accuracy without too much cost.
2. The bottleneck of GhostV2Net can save significant hardware resources while maintaining accuracy.
3. The Wise-IoUv3 loss function can reduce the negative impact of low-quality data in the fire dataset.
4. CBAM can highlight the objects from channel and spatial dimensions and enable the model to focus more on early warning signals of fire accidents.

The remaining sections are organized as follows. The modules, mechanisms, and their functions are explained in Section 2. The dataset and experimental settings are described in Section 3. Then, Section 4 introduces the experimental results and the evaluation metrics. Finally, Section 5 concludes the paper.

2. Method

2.1. YOLOv8

YOLOv8 stands out as an excellent object-detection network. It is well suited for various computer vision tasks, including object detection, instance segmentation, and image classification.

As depicted in Figure 2, YOLOv8 has three parts: backbone, neck, and head. The backbone and neck include C2f, CBS, and SPPF modules. These modules excel at learning key features in images, offering strong feature-representation capabilities. The CBS module is a basic convolution module. The C2f module uses the CSP [22] (cross-stage partial) technique to enhance the learning ability of the model and make the model lighter. The SPPF (spatial pyramid pooling fast) module is a special SPP [23] (spatial pyramid pooling) module. It conducts pooling operations on features with different sizes and dimensions and combines these features into a fixed-size feature map. The parallel structure of the pooling layers in the SPP module is modified into the serial structure in the SPPF module. In addition, SPPF also uses pooling layers with the same kernel size to extract features, effectively reducing the computational cost.

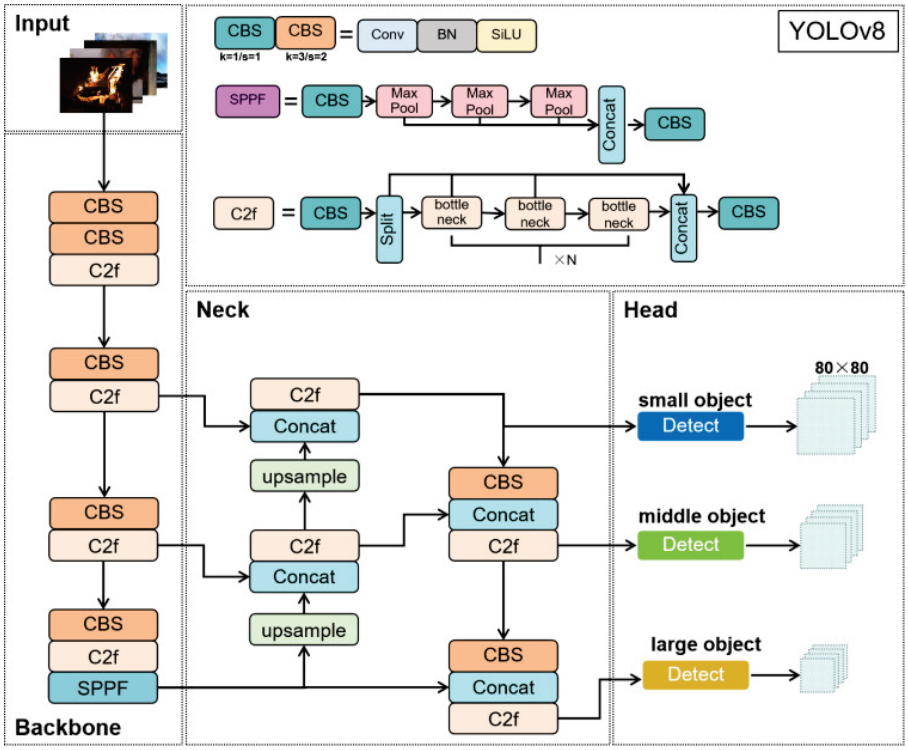


Figure 2. YOLOv8.

The neck is built on the structures of the FPN (feature pyramid network) and PAN (pyramid attention network), which can generate feature maps of different scales containing both semantic and feature information. These two structures enhance the detection capabilities of the model for objects of varying sizes. The head consists of three object-detection layers that perform inference operations on the output from the neck, including locating objects and determining their categories.

2.2. Fire-RPG

YOLOv8 has three detection layers. When the input image size is 640 × 640, the three detection layers can identify large objects above 32 × 32, medium objects above 16 × 16, and small objects above 8 × 8. However, the distance between the camera and the ignition point may not be close in a real urban situation. This means that the flame and

smoke objects whose size is below the smallest detection threshold of YOLOv8 cannot be detected immediately. The photographic distance also affects the detection performance of the model. The farther the distance, the less reliable the network's prediction. Thus, the detection capability for small objects is crucial to provide timely warnings for fire accidents.

To enable the detection network to detect these potential fire hazards early, we integrated an extra detection layer for very small objects above 4×4 into the model. Due to the narrow channel widths, the extra detection layer decreases the parameters. However, the new detection layer enlarges the computational cost, affecting the real-time detection ability, which is the fundamental requirement of fire detection.

Five high-computational-cost C2f modules are replaced with GhostV2C2f modules. This measure significantly reduces the parameters and computational cost. In order to more efficiently extract feature information, the independent CBS modules were replaced with high-performance RepVGG blocks. Moreover, three CBAM blocks were deployed in the backbone. In these positions, CBAM can utilize the abundant spatial and channel information to highlight medium and small objects. The structure of Fire-RPG is illustrated in Figure 3.

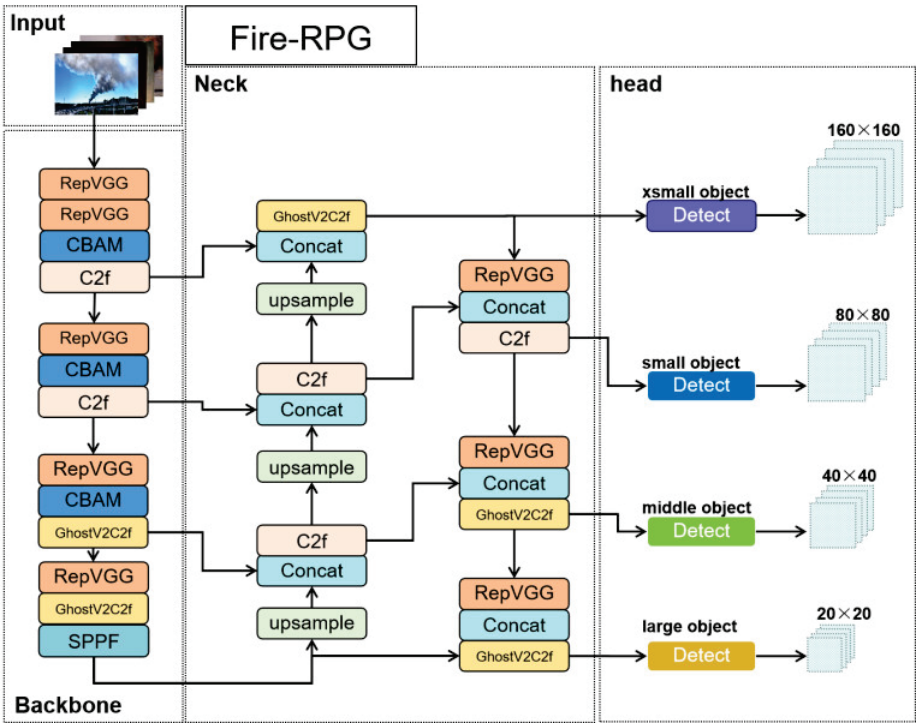


Figure 3. Structure of Fire-RPG.

2.3. RepVGG Block

In order to provide a reliable detection service, current object-detection algorithms typically use many complex structures in their models. Complex convolutional structures have stronger feature learning and representation capabilities compared to simple convolutional structures, so they have higher accuracy. However, complex structures also have significant drawbacks, such as reduced inference speed and hardware unfriendliness.

For fire-warning tasks, complex convolutional structures can improve the accuracy of fire warnings and reduce error rates. However, the reduced inference speed and strict deployment conditions decrease the usefulness of the overall model. The introduction of complex convolutional structures poses a dilemma for fire detection networks. A high-performance and low-cost module is needed to further improve detection performance.

Ding [24] proposed a convolutional structure called RepVGG. This convolutional structure uses the structural reparameterization technique to transform complex network

structures into simple equivalent network structures. During the training phase, RepVGG uses complex structures to participate in training, enabling the model to acquire learning capabilities beyond those of simple structures. In the inference phase, the complex network structures are fused into simple equivalent network structures, reducing the model's inference time.

The structure reparameterization process of RepVGG includes two fusions, as shown in Figure 4.

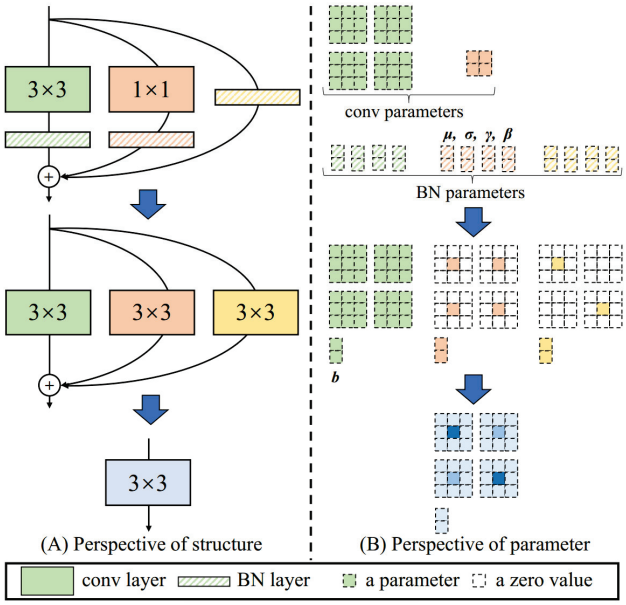


Figure 4. Structural reparameterization process.

The initial state consists of three branches. Each branch has a BN (batch normalization) layer. However, two of the three branches have a 1×1 convolution layer and a 3×3 convolution layer, respectively. All convolution layers have no bias.

The first fusion is that a convolution layer and a BN layer on the same branch fuse into a 3×3 convolution layer with bias. Before the first fusion, a 3×3 convolution layer without bias is needed on each branch. Therefore, some adjustments are necessary on the second branch with a 1×1 convolution layer and the third branch without any convolution layer.

For the second branch, the 1×1 convolution layer can be converted into a 3×3 convolution layer by padding its kernel with 0. The third branch constructs a 3×3 convolution, with the values of its convolutional kernel shown in Figure 5.

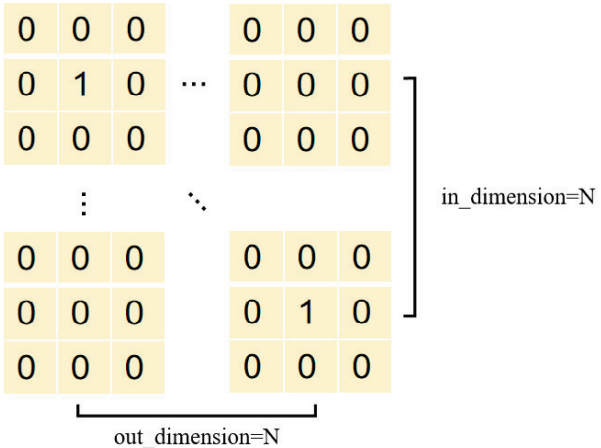


Figure 5. The convolutional kernel in the third branch.

The parameters of a BN layer include μ, σ^2, β, and γ. A feature map x is inputted into the BN layer to produce a new feature map y using the following computation. A small constant ε is introduced into the formula to prevent division by zero.

$$y = \gamma \cdot \left(\frac{x - \mu}{\sqrt{\sigma^2 + \varepsilon}} \right) + \beta \tag{1}$$

The convolution layer without bias only has kernel parameters W. The parameters of a BN layer can be converted into a set of weight parameters and bias parameters.

The weight parameters from the BN layer are multiplied with W to generate the fused convolution kernel parameters W_{fusion}. The bias parameters from the BN layer can serve as the fused convolution bias parameters B_{fusion}.

$$W_{fusion} = \frac{\gamma}{\sigma} \cdot W \tag{2}$$

$$B_{fusion} = \beta - \frac{\mu\gamma}{\sigma} \tag{3}$$

Now, each branch has a 3×3 convolution layer with bias. The final result y is the sum of the results of these convolution layers. Thus, a feature map x needs to go through all branches. The calculation process is as follows:

$$\begin{aligned} y &= W_1 \otimes x + B_1 + W_2 \otimes x + B_2 + W_3 \otimes x + B_3 \\ &= (W_1 + W_2 + W_3) \otimes x + (B_1 + B_2 + B_3) \end{aligned} \tag{4}$$

The second fusion is the process of fusing three branches into a 3×3 convolution layer with bias. This 3×3 convolution layer is the inference structure of RepVGG, with its kernel parameters $W_{inference}$ and bias parameters $B_{inference}$ as follows:

$$W_{inference} = W_1 + W_2 + W_3 \tag{5}$$

$$B_{inference} = B_1 + B_2 + B_3 \tag{6}$$

The RepVGG block is a high-performance convolutional module without a high computational cost. It satisfies the requirements for fire detection. Thus, we replaced the CBS modules with RepVGG blocks, as shown in Figure 6.

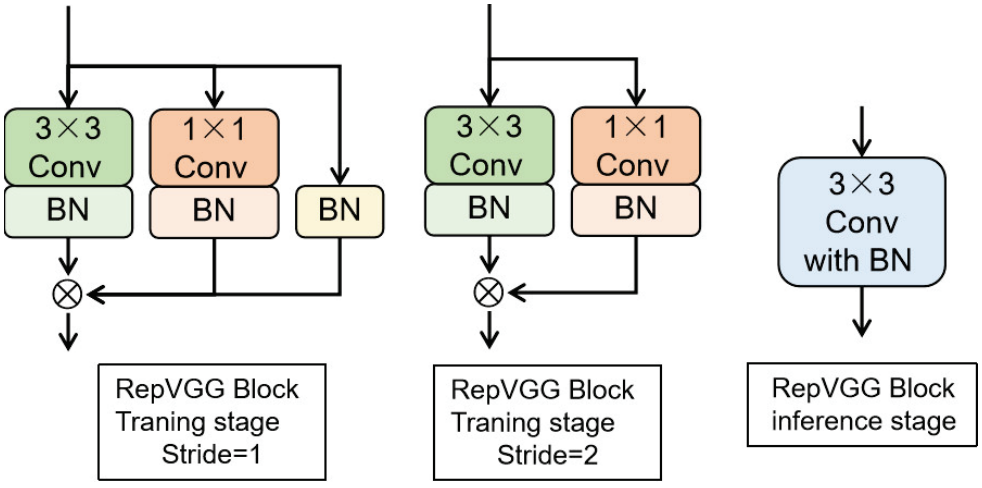


Figure 6. RepVGG block.

2.4. GhostV2 Bottleneck

The extra detection layer significantly improved the detection performance but also increased the parameters and inference time. Lightweight convolutional structures can reduce the parameters and computational cost, enhancing the deployment capabilities of the model. However, an issue that cannot be ignored is that lightweight convolutional structures may reduce the accuracy of model.

Tang [25] optimized the bottleneck structure of GhostNet and introduced the DFC (Decoupled Fully Connected) attention mechanism. They proposed the latest version of GhostNet, GhostV2Net. The bottleneck in GhostV2Net has the advantages of low computational overheads and low parameters and maintains high accuracy.

The GhostV2 bottleneck consists of two components: the DFC attention mechanism and the Ghost module, as shown in Figure 7.

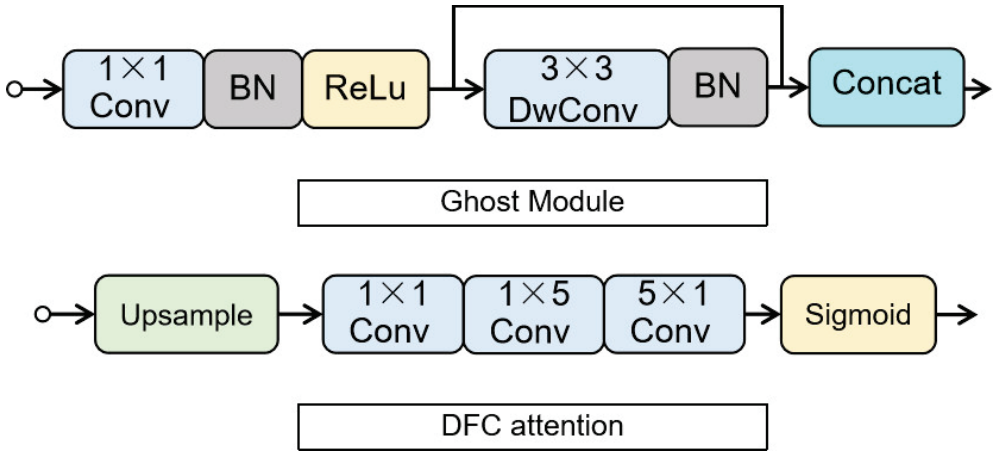


Figure 7. The DFC attention mechanism and Ghost module.

The Ghost module produces feature maps through its cheap operations, as shown in Figure 8. A standard convolution generates only a small part of the feature maps, called intrinsic feature maps. Then, a depth-wise convolution uses the intrinsic feature maps to generate the rest of the feature maps, called redundant feature maps. The final output of the Ghost module is the concatenation of the intrinsic feature maps and the redundant feature maps.

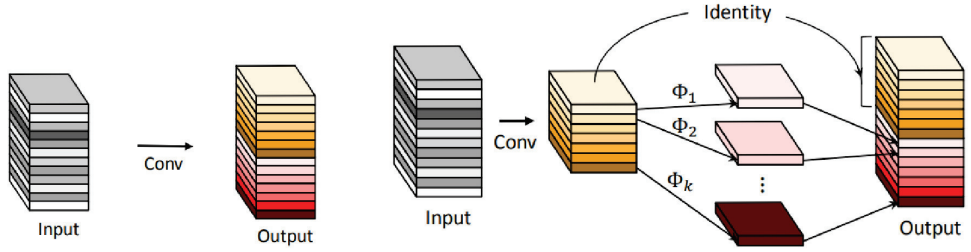


Figure 8. Cheap operation of the Ghost module.

Tang et al. found that fully connected layers can generate attention maps with a global receptive field. Compared to self-attention mechanisms, fully connected layers are easier to implement and have lower computational costs. They also discovered that decoupling the fully connected layers into horizontal and vertical directions can further reduce computational complexity while preserving the completeness of the spatial information. Therefore, the researchers implemented the DFC attention mechanism by using two decoupled, fully

connected layers. The DFC attention mechanism has the advantages of a simple structure, low computational complexity, and hardware friendliness.

Although the cheap operations allow the Ghost module to significantly improve computational efficiency, these operations inevitably weaken the representation capabilities of the Ghost module due to the loss of spatial information.

Therefore, the DFC attention mechanism has been introduced to enhance the output features of the Ghost module. Unlike typical attention mechanisms, the DFC attention mechanism and the Ghost module use the same input and operate in parallel branches. To reduce the burden of parallel computation, the DFC attention mechanism employs downsampling in both vertical and horizontal directions, compressing the feature map to 25% of the original size.

The bottleneck structure of GhostV2Net uses the inverted bottleneck structure. The first Ghost module is used to expand features and increase the number of channels, while the second module reduces the number of channels. This inverted bottleneck naturally decouples the parameters and computational cost from representational capabilities. The structure of GhostV2C2f is shown in Figure 9.

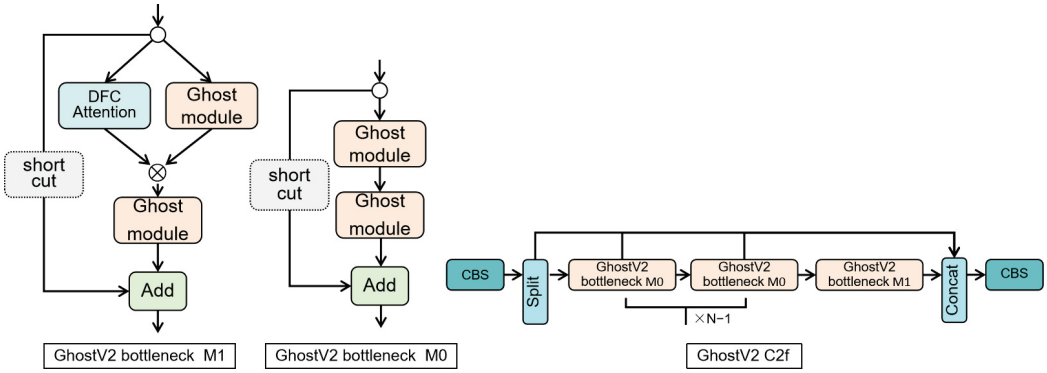


Figure 9. Structure of GhostV2C2f.

While the C2f module can effectively extract features from images, it brings about problems of high computational cost and slow inference speed when the model depth is increased or its width is expanded.

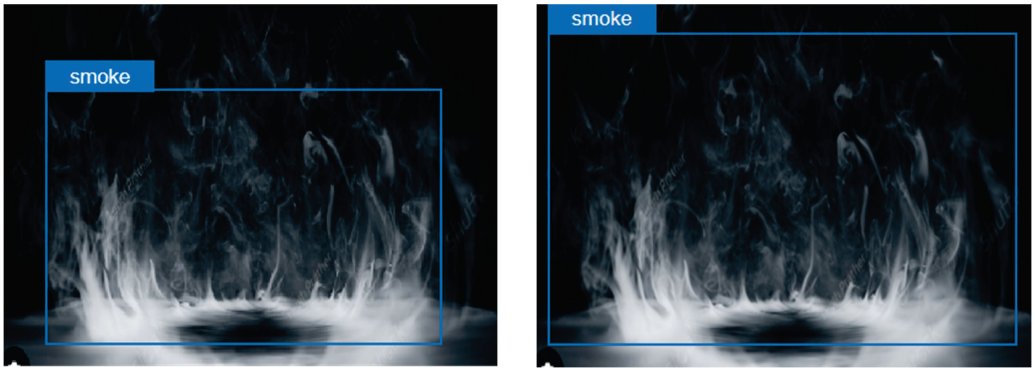
We selected five high-cost C2f modules with large parameters. These high-cost C2f modules were replaced with Ghostv2 C2f modules, containing Ghostv2 bottlenecks in their structure.

To avoid the overfitting caused by the continuous use of attention mechanisms, two types of Ghostv2 bottlenecks were used in the Ghostv2 C2f modules.

As a result, the Ghostv2 C2f modules decreased the parameters and computational cost while maintaining high accuracy. The introduction of Ghostv2 bottlenecks significantly reduced the hardware resources required by the fire-detection model, providing ample space for further improvements.

2.5. Wise-IoUv3

Annotation work is generally carried out manually, so the annotation principles of different annotators can affect the distribution of data in the dataset. It is inevitable that annotators will include large amounts of irrelevant background information or miss some parts of the objects. This issue is particularly obvious in fire datasets because there are often blurred boundaries between objects and the background in real-world fire accident images. For example, the boundaries between thin smoke and the background are often indistinct. A single object may have different annotations, as shown in Figure 10.



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

Figure 10. Two annotation principles. (**a**) The bounding box encloses the main part of the object, but misses the other parts; (**b**) the bounding box encloses the whole object, but also includes too much irrelevant information.

Regardless of whether the data omit some object information or include too much irrelevant background information, these low-quality data can have a detrimental effect on the model.

Many existing studies on IoU loss functions focus on fitting capabilities, like EIoU and SIoU. However, IoU loss functions with high fitting capabilities may decrease the model's accuracy rather than improve it when the dataset quality is unreliable. Therefore, fire-detection tasks require an IoU loss function that effectively reduces the harmful impact of low-quality data.

Wise-IoU [26] (WIoU) can reduce the impact of high-quality and low-quality anchor boxes, allowing the model to focus on normal-quality anchor boxes. WIoU has three versions. WIoUv1 uses distance attention and IoU loss function, while WIoUv2 and WIoUv3 are based on WIoUv1 but apply different kinds of outlier-degree-focusing mechanisms. The WIoUv3 loss function is shown as follows:

$$L_{WIoUv3} = L_{WIoUv1} \cdot r \tag{7}$$

The first factor of WIoUv3 is the WIoUv1 loss function L_{WIoUv1}. It includes the IoU loss function L_{IoU} and the distance attention mechanism R_{WIoU}. R_{WIoU} can increase the IoU loss for normal-quality anchor boxes, while L_{IoU} can reduce the distance attention of high-quality anchor boxes.

The formulation of WIoUv1 is as follows:

$$L_{WIoUv1} = L_{IoU} \cdot R_{WIoU} \tag{8}$$

$$L_{IoU} = 1 - IoU \tag{9}$$

$$R_{WIoU} = exp\left(\frac{(x - x_{gt})^2 + (y - y_{gt})^2}{W_g^2 + H_g^2}\right) \tag{10}$$

In the distance attention mechanism R_{WIoU}, x and y stand for the central coordinates of the anchor box, whereas x_{gt} and y_{gt} stand for the central coordinates of the ground truth bounding box.

W_g and H_g represent the width and height of the smallest enclosing box encompassing both the anchor box and the ground truth bounding box. W_g and H_g are detached from the computational graph to prevent intervening in convergence.

The second factor of WIoUv3 is the non-monotonic gradient gain coefficient r. It is determined by the outlier degree β and hyperparameters α and δ. The formula for r is as follows:

$$r = \frac{\beta}{\delta \alpha^{\beta - \delta}} \tag{11}$$

The outlier degree β is equal to the ratio of L^*_{IoU} to $\overline{L}_{IoU}$. L^*_{IoU} represents the IoU loss detached from the computational graph, while $\overline{L}_{IoU}$ represents the exponential running average of IoU loss with momentum. $\overline{L}_{IoU}$ updates after the new IoU loss of each batch is calculated. The formula for β is as follows:

$$\beta = \frac{L^*_{IoU}}{\overline{L}_{IoU}} \tag{12}$$

Low-quality anchor boxes have a high IoU loss. Therefore, a high outlier degree is assigned to low-quality anchor boxes, while high-quality anchor boxes obtain a low outlier degree. The gradient gain coefficient r can serve as a penalty term to adjust the loss function value for anchor boxes of different qualities.

Our dataset also had issues with low-quality data. Therefore, we decided to use the WIoUv3 loss function to replace the CIoU loss function. By setting the hyperparameters α and δ to 1.9 and 3, respectively, we achieved the final results, as shown in Figure 11.

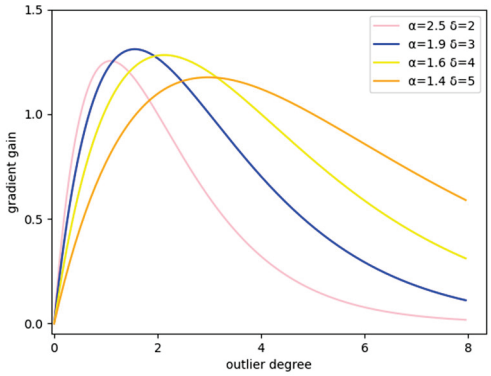


Figure 11. Relationship between gradient gain and outlier degree.

2.6. CBAM

Attention mechanisms play an important role in enhancing the learning capabilities of models. They highlight key features in the original feature maps and ultimately generate attention maps. These attention maps are usually multiplied with the original feature maps as the weight factor of feature information.

In fire scenes, small objects can be enhanced through attention mechanisms. However, each attention mechanism works in a different manner. For instance, the SE [27] (Squeeze-and-Excitation) attention mechanism focuses on channel information, using fully connected layers to generate attention maps. However, SE attention does not consider spatial information, limiting the performance of the SE. The CA [28] (Coordinate Attention) mechanism takes both the channel information and the spatial information into account. However, its computational process is complex and requires permuting operations on the feature maps, which actually prolongs the inference time and significantly reduces the network detection speed at high resolutions.

CBAM (Convolutional Block Attention Module) [29] has a simple structure, low computational cost, and plug-and-play characteristics. It can be seamlessly and cost-effectively integrated into various CNN models. It captures the interdependencies between the feature information in fire images.

CBAM sequentially applies channel and spatial attention mechanisms. The channel attention focuses on "what" is meaningful, while the spatial attention mechanism focuses on

"where" a meaningful part is located. These two attention mechanisms are complementary to each other. The structure of CBAM is shown in Figure 12.

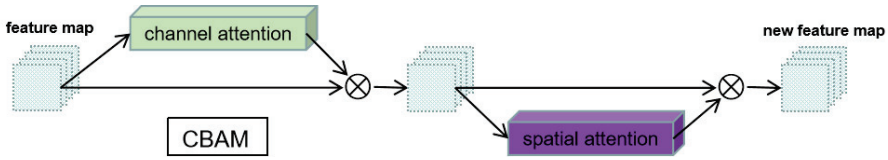


Figure 12. Structure of CBAM.

The channel attention mechanism uses average pooling and max pooling layers and squeezes feature maps of dimensions $C \times H \times W$ into a feature vector of $C \times 1 \times 1$. Each pooling layer produces its vector independently. A shared network extracts channel information from these vectors to generate the corresponding results. The channel attention maps are acquired after applying the sigmoid function on the sum of all pooling results. Similarly, spatial attention also squeezes the feature map with two pooling layers, generating two feature vectors of $1 \times H \times W$. Two vectors are concatenated and inputted into a convolution layer to produce the spatial attention maps. These maps also need to be processed by the sigmoid function. The attention mechanisms in CBAM are shown in Figure 13.

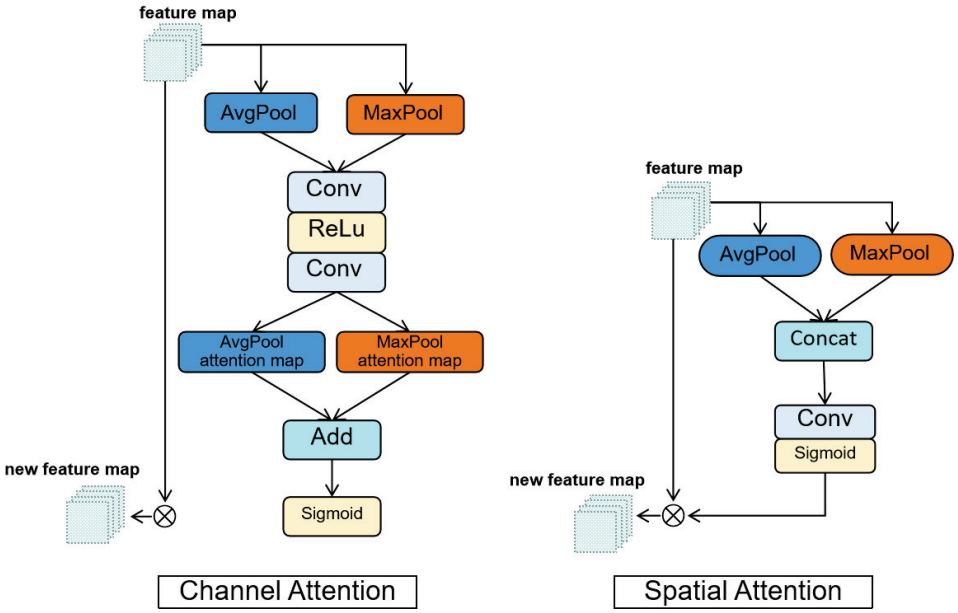


Figure 13. Attention mechanisms in CBAM.

Although many powerful and novel attention mechanisms have been proposed, we found that CBAM is still effective. This attention mechanism can generate dual-dimensional attention maps at a low cost. Therefore, we integrated CBAMs into the model to enhance the learning capabilities for medium and small objects.

3. Dataset and Experimental Settings

We established a dataset of 7052 images. There were two sources of images in our dataset: fire videos and pictures from the Internet. The frames were extracted from the fire accident videos that recorded the entire process from fire accident occurrence to escalation. These fire video frames contained the features of early stage fire accidents and formed a major part of our dataset. In order to ensure the generalization ability of the model, we also randomly obtained images from the Internet by using a web crawler. Several images from the dataset are depicted in Figure 14.

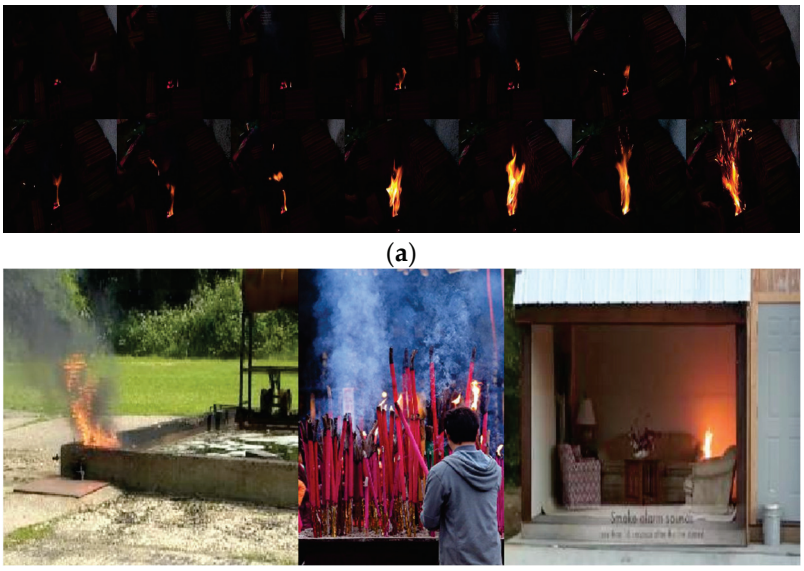


Figure 14. Examples in the dataset. (**a**) Frames extracted from the videos; (**b**) images obtained from the Internet.

Labeling software was utilized to annotate the categories of objects in the images: fire and smoke. The labels were saved in txt format. The dataset was split randomly into training, validation, and test sets at an 8:1:1 ratio, as shown in Table 1.

Table 1. Dataset information.

Dataset	Train	Val	Test	Total
Image	5642	705	705	7052

According to the severity of the fire accident in the image, fire objects were classified into faint flames and strong flames. Faint flames are usually small-sized objects, but in some images taken from a close distance, faint flames appear as medium-sized objects. A similar issue also occurs in the image of strong flames due to the photographic distance.

Smoke objects are classified according to the degree to which they blur the background. Smoke that completely obscures the background is classified as dense smoke, while smoke that allows some background information to be seen is classified as thin smoke. Due to the diffusion of smoke and other factors within the scene, smoke objects are usually medium-sized or large-sized.

We collected statistics on the dataset for the numbers of faint flames, strong flames, thin smoke, and dense smoke, as shown in Table 2.

Table 2. Object information.

Object	Train	Val	Test	Total
Faint flame	5424	750	654	6828
Strong flame	3599	362	323	4284
Thin smoke	1436	192	141	1769
Dense smoke	3437	424	468	4329

The bounding box information in the dataset was visualized. The distributions of the center and sizes of the bounding boxes are displayed in Figure 15.

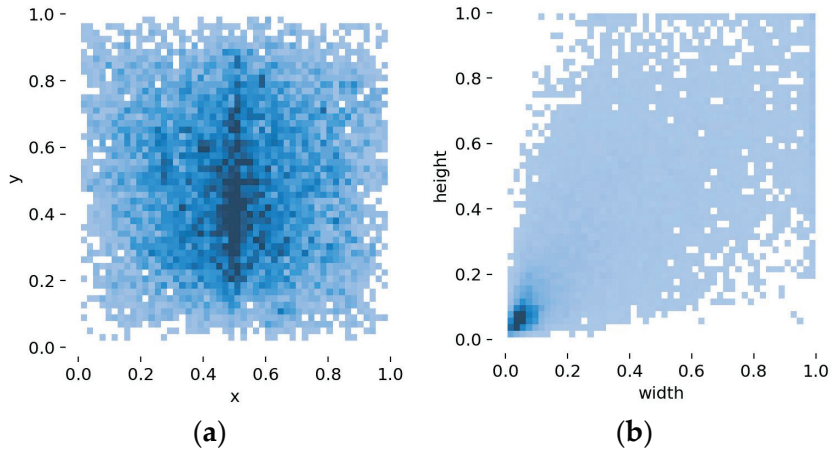


Figure 15. Distribution of bounding box information. (**a**) Distribution of bounding box center position; (**b**) distribution of bounding box size.

In Figure 15a, the x and y coordinates represent the center position of the bounding box relative to the bottom left corner of the image, where (0,0) represents the bottom left corner of the image and (0.5,0.5) represents the center of the image. Likewise, (1,1) represents the top right corner of the image. From this, it is evident that the centers of the bounding boxes tend to cluster in the middle of the image, which corresponds to the reality that photographers usually center their lenses on an object. In Figure 15b, the height and width indicate the size of the bounding box relative to the image. The distribution of the bounding box sizes is concentrated in the lower left corner of the coordinate system, while the rest is spread evenly throughout. This indicates that a large number of objects are small, and the rest of the objects are different in size. Overall, the objects in our dataset are more concentrated in the middle of the image, and most are small objects, which aligns with the natural situation of urban early stage fire accidents.

The experimental environment and hyperparameters are displayed in Tables 3 and 4 to facilitate the replication of our experimental results.

Table 3. Environment settings.

Operating System	CPU	GPU	Programming Language	Deep Learning Framework
Windows 10	Intel i7-13700KF	RTX 3070Ti	Python 3.11	PyTorch 2.0.1

Table 4. Hyperparameter settings.

Batch	Epochs	Patience	Image Size	Optimizer
8	300	50	640	SGD

4. Results and Evaluation

This section provides a detailed introduction to the evaluation metrics used and the experiments implemented in our study.

4.1. Evaluation Metrics

In this study, we utilize classical object-detection metrics, including precision (P), recall (R), mean average precision (mAP), floating-point operations (FLOPs), and parameters. The first three metrics were used to evaluate the detection performance of the model, while FLOPs and parameters assessed the time and space complexity, respectively.

Precision measures the accuracy of correctly predicted positive samples out of all positive samples. TP (true-positive) refers to the samples correctly predicted to be positive,

while FP (false-positive) indicates the samples incorrectly predicted to be positive. The number of all positive samples is the sum of *TP* and *FP*. The formula for precision is as follows:

$$P = \frac{TP}{TP + FP} \tag{13}$$

Recall measures the accuracy of correctly predicted positive samples out of all true samples. *FN* (false-negative) are the samples incorrectly predicted as negative, so FNs are actually true samples. All true samples are the sum of *TP* and *FN*, and the formula for recall is as follows:

$$R = \frac{TP}{TP + FN} \tag{14}$$

AP (average precision) is a metric that measures the detection performance for a specific category. A PR curve (precision–recall curve) can be plotted by using precision and recall values across various confidence thresholds. *AP* is equal to the area underneath the *PR* curve. The formula for *AP* is as follows:

$$AP = \int_0^1 P(R)DR \tag{15}$$

mAP is the mean value of the *AP* for all categories in the dataset. When the number of categories is N, the formula for *mAP* is as follows:

$$mAP = \sum_{i=1}^{N} AP_i \tag{16}$$

In actual experiments, mAP50 and mAP50-95 are usually calculated to assess the comprehensive detection performance of the model. mAP50 is the *mAP* value with an IoU threshold of 50%, commonly used as *mAP*. The IoU threshold is utilized to judge the accuracy of the localization. If the IoU score of a predicted bounding box exceeds the IoU threshold, the localization of the bounding box could be seen as accurate localization.

mAP50-95 calculates the average of the *mAP* values at the IoU thresholds ranging from 50% to 95% with a stride of 5%, providing a broader assessment of the object-detection algorithm performance.

4.2. Comparison of Models

The hardware devices commonly used in urban environments do not reach the level of the laboratory. However, this does not mean that detection performance has to be sacrificed just for better deployment capabilities. Urban systems can select a model that balances parameters, computational cost, and detection performance well. With this in mind, this study adopted YOLOv8s as the baseline model, rather than a smaller scale of YOLOv8, as shown in Table 5 and Figure 16.

Table 5. Comparison of models.

Model	mAP50	mAP50-95	Parameters	GFLOPs
Faster RCNN	73.0%	46.2%	41755286	134.4
SSD	71.6%	44.3%	35641314	34.9
RetinaNet	77.7%	46.6%	34014999	151.5
YOLOv10s [30]	78.9%	46.1%	8067900	24.8
YOLOv8s	79.1%	47.7%	11136374	28.6
Fire-RPG	81.3%	48.9%	7808446	32.4

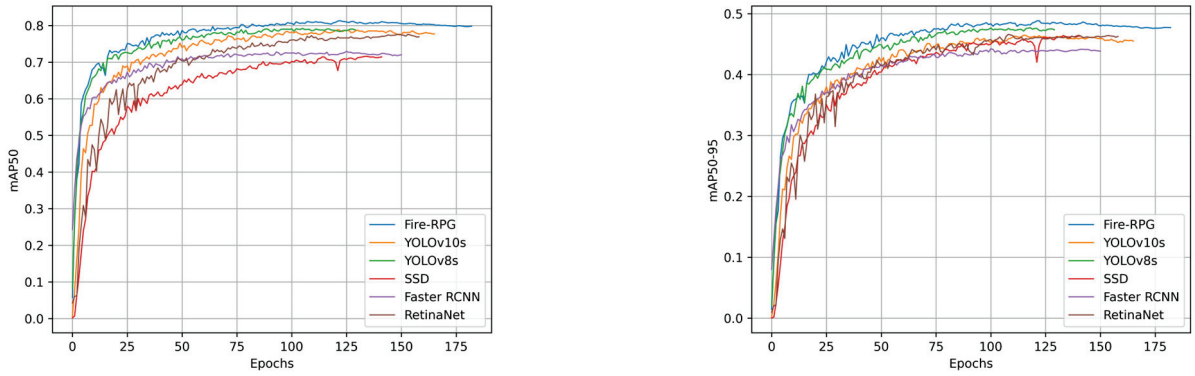


Figure 16. Comparison of mAP: (**a**) mAP50; (**b**) mAP50-95.

We conducted the comparative experiments between Fire-RPG and various mainstream models under the same experimental settings.

Besides YOLOv8s and Fire-RPG, the models involved in the comparative experiments included Faster R-CNN, SSD, RetinaNet, and YOLOv10s. Compared to the latest YOLO series algorithm, YOLOv10, YOLOv8s had higher parameters and computational costs. However, YOLOv8s also achieved higher accuracy. It is evident that Fire-RPG achieved the highest mAP value with relatively low parameters and FLOPs.

4.3. Comparison of IoU Loss Functions

This study used YOLOv8s to compare the influence between WIoUv3 and other loss functions. The change in loss functions does not affect the parameters or FLOPs, so these two metrics are not taken into account.

As shown in Table 6, IoU loss functions with high-fitting capabilities, like SIoU and EIoU, negatively influence detection performance. In contrast, simple but low-fitting IoU loss functions like GIoU and DIoU achieve high mAP scores. This proves that the influence of data quality cannot be ignored. WIoUv3 maintains high detection performance, and its mAP reaches 80.0%, proving that the focus mechanism based on the outlier degree can effectively reduce the harmful impact of low-quality examples.

Table 6. Comparison of IoU loss functions.

Model	P	R	mAP50	mAP50-95
YOLOv8s-CIoU	84.8%	70.9%	79.1%	47.7%
YOLOv8s-GIoU [31]	85.1%	70.4%	79.6%	47.6%
YOLOv8s-DIoU [32]	82.3%	71.6%	79.8%	47.5%
YOLOv8s-SIoU [33]	82.6%	71.4%	79.1%	47.2%
YOLOv8s-EIoU [34]	82.5%	71.9%	79.2%	47.4%
YOLOv8s-WIoUv1	85.1%	70.3%	79.0%	47.0%
YOLOv8s-WIoUv2	84.7%	70.7%	79.6%	47.6%
YOLOv8s-WIoUv3	84.2%	71.4%	80.0%	47.3%

4.4. Comparison of Attention Mechanisms

We also compared the detection performance between CBAM and other attention mechanisms. The attention mechanisms usually function across the channel or spatial dimensions. The same deployment location affects the performance of different attention mechanisms to different degrees, so we deployed different attention mechanisms at different locations. However, their deployment locations were still confined to the backbone in order to keep the conditions as consistent as possible.

As shown in Table 7, The results indicate that CBAM exhibits excellent enhancement in detection performance, with mAP50 of 80.0% and mAP50-95 of 48.0%. Although the

parameter increase for CBAM is also the highest, it is acceptable compared to other non-attention mechanism modules.

Table 7. Comparison of attention mechanisms.

Model	P	R	mAP50	mAP50-95	Parameters	GFLOPs
YOLOv8s	84.8%	70.9%	79.1%	47.7%	11136374	28.6
YOLOv8s-SE	82.7%	71.7%	79.6%	47.9%	11309526	28.7
YOLOv8s-SA [35]	84.6%	70.4%	79.4%	48.0%	11136542	28.7
YOLOv8s-ECA [36]	82.1%	71.1%	79.2%	47.4%	11136383	28.7
YOLOv8s-CA	83.5%	71.5%	79.8%	48.0%	11142270	28.7
YOLOv8s-SimAM [37]	82.0%	71.1%	79.3%	47.8%	11136374	28.6
YOLOv8s-CBAM	83.2%	72.9%	80.0%	48.0%	11223132	28.7

4.5. Ablation Experiment

An ablation experiment was conducted to validate the necessity of each module introduced. The enhancement in Fire-RPG is segmented into five stages. The impact of each module is evaluated by comparing the metrics between the current stage and the preceding stages. The results are shown in Table 8.

Table 8. Results of ablation experiments.

Model	P	R	mAP50	mAP50-95	Parameters	GFLOPs
YOLOv8s	84.8%	70.9%	79.1%	47.7%	11136374	28.6
YOLOv8s-WIoUv3	84.2%	71.4%	80.0%	47.3%	11136374	28.6
YOLOv8s-WIoUv3 -Detect Layer	81.9%	72.2%	80.6%	48.1%	10637496	37.0
YOLOv8s-WIoUv3 -Detect Layer-RepVGG	84.9%	72.9%	80.9%	48.0%	10626968	36.6
YOLOv8s-WIoUv3 -Detect Layer-RepVGG -GhostV2C2f	85.5%	73.2%	80.7%	48.4%	7380504	31.3
Fire-RPG	85.9%	72.1%	81.3%	48.9%	7808446	32.4

In stage one, the CIoU loss function is replaced with WIoUv3. This improvement does not change the parameters or FLOPs. Recall and mAP50 increase by 0.5% and 0.9%, respectively, but precision and mAP50-95 decrease. In stage two, the extra object-detection layer is added, reducing the parameters but increasing the FLOPs. The detection performance metrics rise significantly, except for a 2.3% drop in precision. In stage three, the CBS modules are replaced with the RepVGG blocks, bringing a slight decrease in parameters and FLOPs. Precision rises by 3% but only increases by 0.1% compared to YOLOv8s. Stage four is the introduction of the GhostV2 bottleneck, greatly reducing the parameters and FLOPs. Although mAP50 decreases by 0.2%, mAP50-95 has a 0.4% increase. In stage five, the introduction of CBAM slightly increases the parameters and FLOPs. The mAP50 and mAP50-95 enhance by 0.6% and 0.5%, respectively.

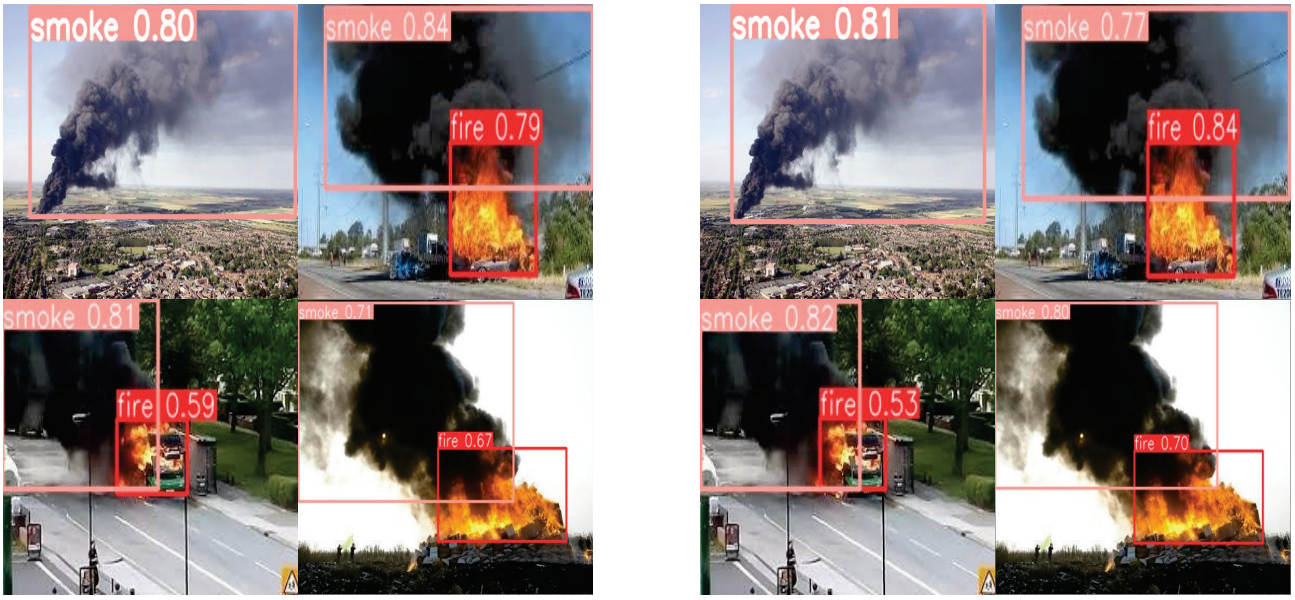
4.6. Detection Results

The detection capabilities of the Fire-RPG model can be explicitly shown by comparing the detection results. The results are shown in Figures 17 and 18.



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

Figure 17. Results of the detection of faint flames and thin smoke. (**a**) YOLOv8; (**b**) Fire-RPG.



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

Figure 18. Results for the detection of strong flames and dense smoke. (**a**) YOLOv8; (**b**) Fire-RPG.

For small and difficult-to-detect objects in images, such as faint flames and thin smoke, Fire-RPG exhibits strong accuracy. It captures extremely small flames and thin smoke sharply in images and assigns them a high confidence value. In contrast, YOLOv8s not only misses these objects but also provides lower confidence scores. This demonstrates that Fire-RPG has better detection capabilities for the early warning signals of fire accidents.

For large-scale and easily detectable objects, such as strong flames and dense smoke, the detection results for Fire-RPG and YOLOv8s are similar. Both models can accurately locate the objects in the images and effectively use bounding boxes to enclose them.

4.7. Detection Performance in Different Scenes

To demonstrate the generalization ability of the model, three extra datasets were used to display the performance of our model in different scenes. We compared the detection performance of YOLOv8s and Fire-RPG to illustrate the effectiveness of our model, as shown in Figure 19.



Figure 19. Detection results for different datasets. (**a**) D-Fire and YOLOv8; (**b**) D-Fire and Fire-RPG; (**c**) ForestFire and YOLOv8; (**d**) ForestFire and Fire-RPG; (**e**) DFS and YOLOv8; (**f**) DFS and Fire-RPG.

The D-Fire dataset focuses on open wilderness scenes, with data primarily sourced from long-term surveillance cameras monitoring the wilderness. This dataset annotates two objects: smoke and flame. To avoid excessive time expenditure, we removed the images without annotation information from this dataset. Following this, the dataset contained 11,689 images.

The ForestFire dataset focuses on forest scenes. To improve the quality of the dataset, the dataset creators used data-augmentation methods such as flipping, rotation, and HSV transformation. This dataset only annotated flame objects, and all images were resized to 640×640.

The DFS dataset contains fire images from various scenes, with all images collected from real scenarios and annotated in a standardized manner. Considering the potential for false-positives caused by objects similar to flames or smoke, the dataset creators annotated these objects as "other." Thus, there are three labels in this dataset: flame, smoke, and other.

We collected the performance indicators under the three datasets and displayed the results in the Table 9.

Table 9. Results of the different scenes.

Dataset	Images	Model	P	R	mAP50	mAP50-95
D-Fire [38]	11689	Fire-RPG	78.7%	71.7%	79.2%	46.1%
		YOLOv8s	76.6%	72.9%	78.0%	45.6%
ForestFire [39]	7763	Fire-RPG	63.2%	60.0%	63.1%	27.8%
		YOLOv8s	63.0%	58.5%	61.8%	27.5%
DFS [40]	9462	Fire-RPG	61.2%	49.2%	54.9%	28.1%
		YOLOv8s	60.7%	48.8%	53.8%	27.3%

The results show that Fire-RPG surpasses YOLOv8s in detection performance in the above datasets. Specifically, in terms of mAP50, Fire-RPG achieved an improvement of over 1% across all three datasets. In terms of mAP50-95, Fire-RPG achieved a 0.8% improvement on the DFS dataset. Hence, Fire-RPG demonstrates high generalization ability and can adapt to various scenarios such as forest fires and wilderness fires.

5. Conclusions

The initial stage of fire accidents includes objects such as faint flames and thin smoke, which are considered early warning signals of fire. Common CNN models fail to provide reliable early warnings due to their poor capacity to detect these signals. We established a fire dataset with numerous images of early stage fire accidents and proposed a fire detection network based on YOLOv8, named Fire-RPG.

In this model, we introduced an extra detection layer for smaller objects, enabling the model to detect smaller flame objects. Next, the lightweight CNN structure, GhostV2 bottleneck, was used to reduce the inference time and optimize the model structure. To address the issue of low-quality data, the Wise-IoUv3 loss function was used during the training phase. Then, we introduced the RepVGG block to improve the network's learning capabilities. Finally, we employed the CBAM attention mechanism to highlight the key objects in feature maps, especially small- and medium-sized smoke fire objects.

In comparative experiments with mainstream CNN models, Fire-RPG achieved the best detection performance, with its mAP of 81.3%, and had low-level parameters and FLOPs. In addition, Fire-RPG showed significant improvement in other fire scenarios compared to YOLOv8s, proving the effectiveness of our improvements. Fire-RPG is suitable as a detection model for urban fire alarm systems. It can utilize a large number of cameras installed in urban systems to capture images and detect faint flames and thin smoke. With more powerful small-object detection capabilities, Fire-RPG can detect fire accidents in their early stages and respond in time. This provides more response time for firefighters and effectively ensures the safety of people's lives and social property.

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

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding authors.

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

References

1. Statistics on Fire and Police Situation across the Country in 2018. Available online: https://www.119.gov.cn/gk/sjtj/2022/54.shtml (accessed on 17 April 2024).
2. In 2019, 233,000 Fires Were Reported Nationwide. Available online: https://www.119.gov.cn/gk/sjtj/2022/386.shtml (accessed on 17 April 2024).
3. National Fire and Police Response Situation in 2020. Available online: https://www.119.gov.cn/gk/sjtj/2022/13721.shtml (accessed on 17 April 2024).
4. Firefighting Calls Hit a New High in 2021, with 745,000 Fires Put Out. Available online: https://www.119.gov.cn/gk/sjtj/2022/26442.shtml (accessed on 17 April 2024).
5. National Police and Fire Situation in 2022. Available online: https://www.119.gov.cn/qmxfxw/xfyw/2023/36210.shtml (accessed on 17 April 2024).
6. Ten Types of Fire Sources That Often Cause Fires. Available online: https://www.gov.cn/ztzl/djfh/content_436344.htm (accessed on 17 April 2024).
7. Gaur, A.; Singh, A.; Kumar, A.; Kulkarni, K.S.; Lala, S.; Kapoor, K.; Srivastava, V.; Kumar, A.; Mukhopadhyay, S.C. Fire sensing technologies: A review. *IEEE Sens. J.* **2019**, *19*, 3191–3202. [CrossRef]
8. Lestari, D.P.; Kosasih, R.; Handhika, T.; Sari, I.; Fahrurozi, A. Fire hotspots detection system on CCTV videos using you only look once (YOLO) method and tiny YOLO model for high buildings evacuation. In Proceedings of the 2019 2nd International Conference of Computer and Informatics Engineering (IC2IE), Banyuwangi, Indonesia, 10–11 September 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 87–92.
9. Ko, B.C.; Cheong, K.H.; Nam, J.Y. Fire detection based on vision sensor and support vector machines. *Fire Saf. J.* **2009**, *44*, 322–329. [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. 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] [PubMed]
12. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Springer International Publishing: Berlin/Heidelberg, Germany, 2016. Proceedings, Part I 14. pp. 21–37.
13. 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, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
14. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
15. Li, P.; Zhao, W. Image fire detection algorithms based on convolutional neural networks. *Case Stud. Therm. Eng.* **2020**, *19*, 100625. [CrossRef]
16. Wang, X.; Cai, L.; Zhou, S.; Jin, Y.; Tang, L.; Zhao, Y. Fire Safety Detection Based on CAGSA-YOLO Network. *Fire* **2023**, *6*, 297. [CrossRef]
17. Lin, J.; Lin, H.; Wang, F. A Semi-Supervised Method for Real-Time Forest Fire Detection Algorithm Based on Adaptively Spatial Feature Fusion. *Forests* **2023**, *14*, 361. [CrossRef]
18. Zhao, L.; Zhi, L.; Zhao, C.; Zheng, W. Fire-YOLO: A Small Target Object Detection Method for Fire Inspection. *Sustainability* **2022**, *14*, 4930. [CrossRef]
19. Zhang, Z.; Tan, L.; Tiong, R.L.K. Ship-Fire Net: An Improved YOLOv8 Algorithm for Ship Fire Detection. *Sensors* **2024**, *24*, 727. [CrossRef] [PubMed]
20. Chen, G.; Cheng, R.; Lin, X.; Jiao, W.; Bai, D.; Lin, H. LMDFS: A Lightweight Model for Detecting Forest Fire Smoke in UAV Images Based on YOLOv7. *Remote Sens.* **2023**, *15*, 3790. [CrossRef]
21. Ultralytics YOLOv8. Available online: https://github.com/ultralytics/ultralytics (accessed on 17 April 2024).
22. Wang, C.H.; Mark Liao, H.Y.; Wu, Y.H.; Chen, P.Y.; Hsieh, J.W.; Yeh, I.H. CSPNet: A new backbone that can enhance learning capabilities of CNN. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 390–391.
23. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef] [PubMed]
24. Ding, X.; Zhang, X.; Ma, N.; Han, J.; Ding, G.; Sun, J. Repvgg: Making vgg-style convnets great again. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 13733–13742.
25. Tang, Y.; Han, K.; Guo, J.; Xu, C.; Xu, C.; Wang, Y. GhostNetv2: Enhance cheap operation with long-range attention. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 9969–9982.
26. Tong, Z.; Chen, Y.; Xu, Z.; Yu, R. Wise-IoU: Bounding box regression loss with dynamic focusing mechanism. *arXiv* **2023**, arXiv:2301.10051.
27. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.

28. Hou, Q.; Zhou, D.; Feng, J. Coordinate attention for efficient mobile network design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 13713–13722.
29. 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.
30. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; Ding, G. YOLOv10: Real-Time End-to-End Object Detection. *arXiv* **2024**, arXiv:2405.14458.
31. Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; Savarese, S. Generalized intersection over union: A metric and a loss for bounding box regression. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 658–666.
32. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU loss: Faster and better learning for bounding box regression. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 12993–13000.
33. Gevorgyan, Z. SIoU loss: More powerful learning for bounding box regression. *arXiv* **2022**, arXiv:2205.12740.
34. Zhang, Y.F.; Ren, W.; Zhang, Z.; Jia, Z.; Wang, L.; Tan, T. Focal and efficient IOU loss for accurate bounding box regression. *Neurocomputing* **2022**, *506*, 146–157. [CrossRef]
35. Zhang, Q.L.; Yang, Y.B. Sa-net: Shuffle attention for deep convolutional neural networks. In Proceedings of the ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 2235–2239.
36. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11534–11542.
37. Yang, L.; Zhang, R.Y.; Li, L.; Xie, X. Simam: A simple, parameter-free attention module for convolutional neural networks. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 18–24 July 2021; pp. 11863–11874.
38. D-Fire. Available online: https://github.com/gaiasd/DFireDataset (accessed on 24 May 2024).
39. ForestFire. Available online: https://universe.roboflow.com/smokedetection-lfdtr/forestfire-aepov (accessed on 24 May 2024).
40. DFS. Available online: https://github.com/siyuanwu/DFS-FIRE-SMOKE-Dataset (accessed on 24 May 2024).

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

Robust Landslide Recognition Using UAV Datasets: A Case Study in Baihetan Reservoir

Zhi-Hai Li [1], An-Chi Shi [1,2], Huai-Xian Xiao [3], Zi-Hao Niu [1], Nan Jiang [3,*], Hai-Bo Li [4] and Yu-Xiang Hu [4]

[1] Power China Huadong Engineering Corporation Limited, Hangzhou 311122, China; li_zh5@hdec.com (Z.-H.L.); shi_ac@hdec.com (A.-C.S.); niu_zh1@hdec.com (Z.-H.N.)
[2] Zhejiang Huadong Geotechnical Investigation & Design Institute Corporation Limited, Hangzhou 310004, China
[3] College of Water Resource and Hydropower, Sichuan University, Chengdu 610065, China; xiaohuaixian@stu.scu.edu.cn
[4] State Key Laboratory of Hydraulics and Mountain River Engineering, Sichuan University, Chengdu 610065, China; hbli@scu.edu.cn (H.-B.L.); yxhu@scu.edu.cn (Y.-X.H.)
* Correspondence: 2018223060073@stu.scu.edu.cn

Abstract: The task of landslide recognition focuses on extracting the location and extent of landslides over large areas, providing ample data support for subsequent landslide research. This study explores the use of UAV and deep learning technologies to achieve robust landslide recognition in a more rational, simpler, and faster manner. Specifically, the widely successful DeepLabV3+ model was used as a blueprint and a dual-encoder design was introduced to reconstruct a novel semantic segmentation model consisting of Encoder1, Encoder2, Mixer and Decoder modules. This model, named DeepLab for Landslide (DeepLab4LS), considers topographic information as a supplement to DeepLabV3+, and is expected to improve the efficiency of landslide recognition by extracting shape information from relative elevation, slope, and hillshade. Additionally, a novel loss function term—Positive Enhanced loss (PE loss)—was incorporated into the training of DeepLab4LS, significantly enhancing its ability to understand positive samples. DeepLab4LS was then applied to a UAV dataset of Baihetan reservoir, where comparative tests demonstrated its high performance in landslide recognition tasks. We found that DeepLab4LS has a stronger inference capability for landslides with less distinct boundary information, and delineates landslide boundaries more precisely. More specifically, in terms of evaluation metrics, DeepLab4LS achieved a mean intersection over union (mIoU) of 76.0% on the validation set, which is a substantial 5.5 percentage point improvement over DeepLabV3+. Moreover, the study also validated the rationale behind the dual-encoder design and the introduction of PE loss through ablation experiments. Overall, this research presents a robust semantic segmentation model for landslide recognition that considers both optical and topographic semantics of landslides, emulating the recognition pathways of human experts, and is highly suitable for landslide recognition based on UAV datasets.

Keywords: landslide recognition; UAV; deep learning; semantic segmentation

1. Introduction

Landslides are one of the most common geological disasters in the world. When a landslide occurs, it often carries a large amount of soil, rock, or a mixture of both, moving downhill along the sliding surface formed on the slope [1]. As a result, the built environment (including houses, factories, roads, etc.) and the natural environment (including farms, forests, grasslands, etc.) on the slope are partially or completely destroyed [2,3]. In some extreme cases, landslides and their secondary disasters can cause considerable damage. For example, the 2014 Abe Barek landslide in Badakhshan Province, Afghanistan, resulted in nearly 2700 fatalities [4]; the 2018 Baige landslide in Tibet, China, blocked the main stream of the Jinsha River and formed a huge barrier lake, which breached, causing

floods that affected areas as far downstream as Shangri-La and Lijiang, nearly 600 km away, resulting in losses exceeding USD 500 million in Lijiang alone [5,6]. However, landslide disasters are not entirely uncontrollable. Effective monitoring of landslide-prone areas and research into landslide initiation mechanisms, motion processes, and risk management, based on monitoring data, can lead to a comprehensive understanding of landslides. This, in turn, allows for the implementation of early measures to reduce the losses from landslide disasters and even prevent them through certain slope reinforcement measures [7–9].

In practice, the effective monitoring of landslide areas is not an easy task. After a major landslide disaster, detailed field investigations, sampling, and experimental work are typically required [10]. However, the mountainous distribution characteristics of landslides often make it difficult for researchers and monitoring equipment to quickly reach the disaster site, thus limiting the possibility of immediate landslide research post-disaster. Fortunately, the recent popularization of low-cost unmanned aerial vehicles (UAV) has greatly facilitated the monitoring of landslides. UAV, with their convenience, speed, proximity, and lightweight dataset [11], perfectly meet the needs for landslide monitoring. The use of UAV photogrammetry technology allows researchers to easily obtain extensive landslide photographs from a distance. Then, combined with structure-from-motion (SfM) technology, very high resolution (VHR) landslide image datasets can be quickly constructed, which can be used as the basic data for landslide identification and monitoring [12].

Following the deformation process of the landslide, a number of physical changes can be observed through photogrammetry. These include alterations in topography [13], vegetation cover [14], color (spectrum) [15], and cracks [16]. Therefore, based on the aforementioned multifaceted information, landslide recognition can be indirectly achieved from the landslide image dataset. Landslide recognition based on the image dataset can be divided into four methods: manual interpolation method, object-oriented method, object-based deep learning (DL) method, and pixel-based DL method. Figure 1 provides an intuitive illustration of the results of these four landslide recognition methods. The manual interpolation method (Figure 1a) employs direct observation of landslide features in images [17], or alternatively, the selection of a threshold through trial and error, followed by semi-automated recognition of landslide areas in images based on this threshold [18,19]. The object-oriented method (Figure 1b) often employs the multiresolution segmentation approach, which first divides the image into multiple non-overlapping objects based on homogeneity by selecting the optimal segmentation scale, and then extracts the landslide objects by combining the spectral, textural, and morphological characteristics of landslides [20,21]. The object-based deep learning (DL) method (Figure 1c) uses algorithms to detect objects in images by determining bounding boxes and classes, either in one or two stages. The YOLO series models are key examples of this method [22,23]. The pixel-based method (Figure 1d) relates to semantic segmentation tasks, which classify each pixel in the image using deep learning techniques, considering both local and regional context [24,25]. Overall, the object-based methods focused on object detection and the pixel-based methods focused on semantic segmentation have come together as the dominant methods for landslide recognition [26,27].

In contrast to object detection, semantic segmentation could provide more physical information about landslides, including boundaries and areas. While current popular semantic segmentation network models such as U-Net [28], Mask R-CNN [29], and DeepLabV3+ [30] can accurately perform landslide recognition tasks, they encounter significant difficulties in complex scenarios. Since these popular models usually use only RGB optical images as inputs, they may experience interference from factors such as vegetation, shading, and human activities in real landslide identification scenarios [31–33]. To address this problem, we hypothesize that the introduction of a corresponding DSM obtained from SfM may be able to ameliorate this problem [34]. Since the DSM essentially represents a 2.5D reproduction of a real-world 3D model, its introduction into the semantic segmentation model can be interpreted as improving the ability to identify landslides from

an engineering geological perspective, in a manner similar to the empirical judgement of human experts during landslide field investigations.

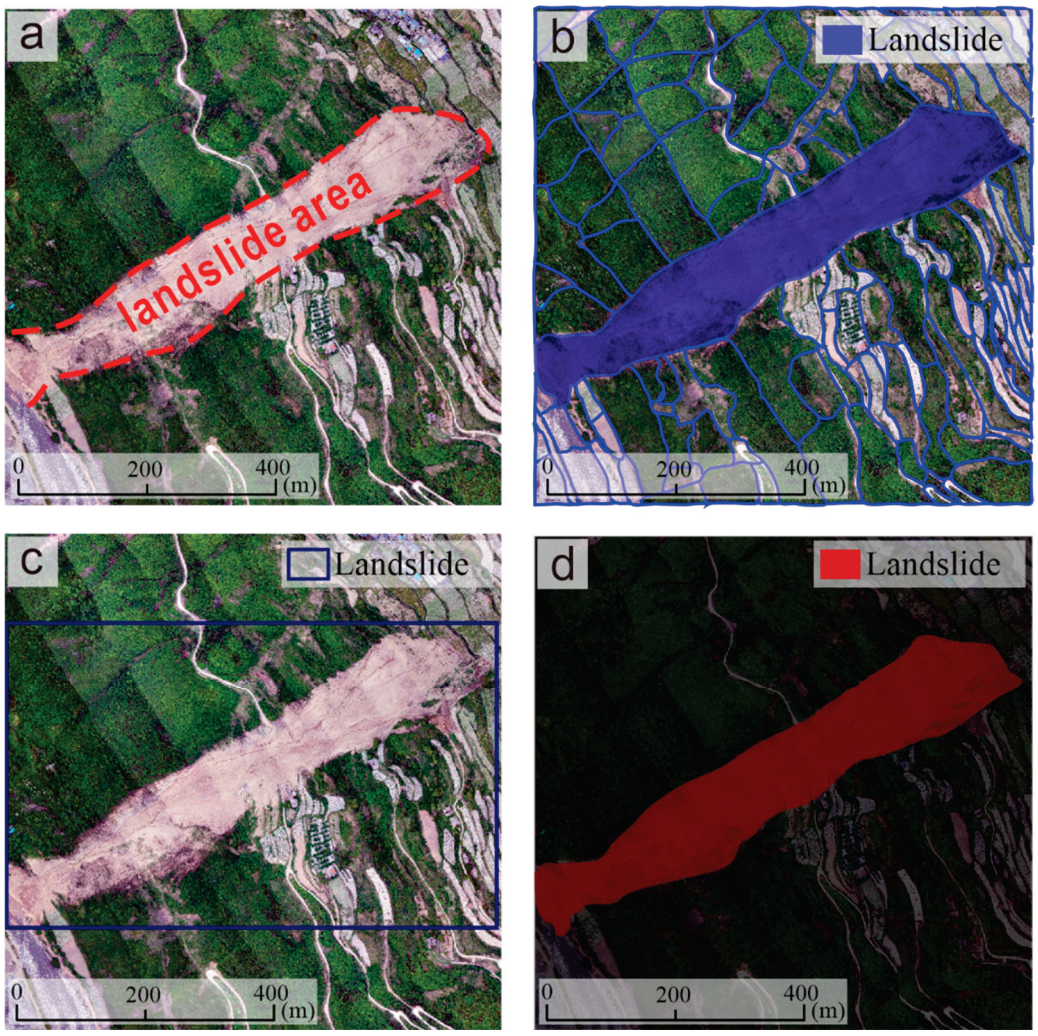


Figure 1. Different landslide recognition methods and their general result based on UAV orthophoto: (**a**) landslide boundary identification by manual judgement; (**b**) landslide identification based on object-oriented method; (**c**) landslide positioning based on object-based DL method; (**d**) landslide identification based on pixel-based DL method.

This study will reference the architecture of DeepLabV3+ to re-design and develop a new semantic segmentation model named DeepLab for Landslide (DeepLab4LS) for landslide recognition that integrates optical and topographic information, which will be validated on landslide UAV datasets from the Baihetan reservoir. Additionally, this paper will investigate the issue of the loss function during model training to determine the most suitable loss function scheme for landslide recognition through detailed experiments. Finally, comparison and ablation experiments will be conducted on the novel semantic model to explore the role of various details in the architecture and training of the semantic model in landslide recognition.

2. Study Area and Dataset

2.1. Study Area and Data Acquisition

The Baihetan reservoir was formed after the Baihetan hydropower station began storing water in April 2021, with a total storage capacity of 20.627 billion m^3 [35]. The storage of water has changed conditions such as the groundwater level and pore water

pressure on the slopes [36], leading to the continuous development of a series of landslides within the reservoir area [16,37,38]. The activity of the landslides has affected the production and traffic of towns along the river within the reservoir area, and the potential impulse wave risk of the landslides also poses a significant threat to the downstream Baihetan hydropower station. In this study, a UAV image dataset obtained from the Baihetan reservoir in the lower reaches of the Jinsha river in China is employed as the data source for model training and validation.

Therefore, to thoroughly investigate the distribution of landslides within the reservoir area, a VHR UAV image dataset was collected between February and September 2023. This dataset, which covers an area of approximately 65 km^2, encompasses a stretch of Jinsha river within the reservoir area measuring approximately 50 km in length (Figure 2). The UAV photogrammetry employed the Feima D2000 four-rotor UAV, equipped with a SONY DOP 3000 camera (Shenzhen Feima Robotics Co., Ltd., Shenzhen, China) (Figure 3). The UAV flew in a fixed flight height mode at an altitude of approximately 150 to 200 m above the ground, ensuring that the resolution of the collected photos was stable at approximately 0.03 m. The heading overlap and the side overlap were 80% and 65%, respectively. In total, data collection for the UAV photogrammetry took approximately 1 month.

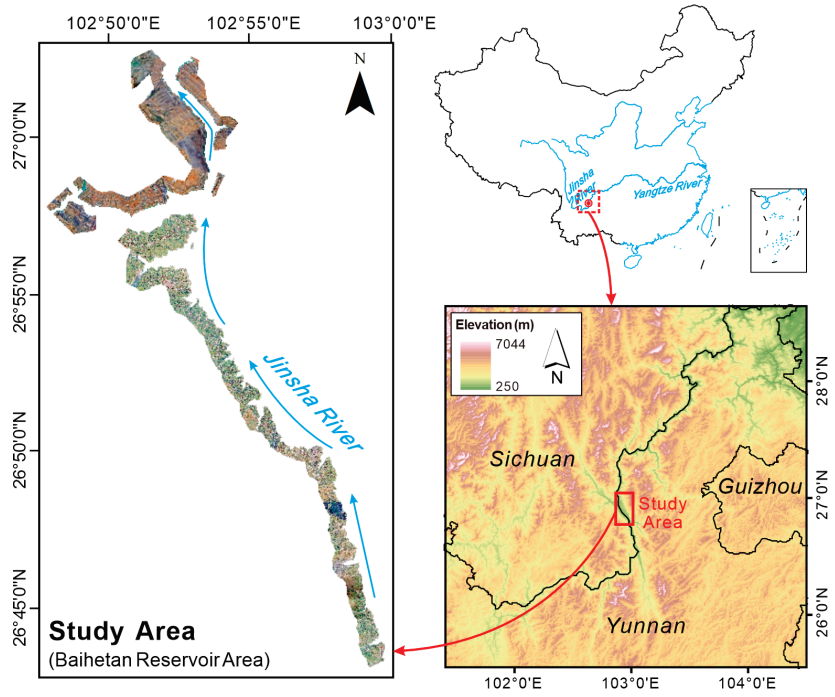


Figure 2. The UAV orthophoto (**left**) and the geographic location (**right**) of the study area. The UAV orthophoto is approximately 50 km in length along the Jinsha River and covers an area of approximately 65 km^2.



Figure 3. UAV equipment: (**a**) photograph of UAV working onsite; (**b**) Feima D2000 UAV; (**c**) SONY-DOP 3000 camera.

The original photos captured by the UAV and the corresponding ground control points (GCPs) were input into Pix4D software (4.9.0 trial version), where they underwent the SfM

process. As a result, the orthophoto (Figure 2) and DSM of the study area were generated. It is worth mentioning that the orthophoto and DSM were subsequently downsampled to 0.1 m to avoid potential issues during subsequent deep learning training. At a resolution of 0.03 m, a 1024 × 1024 size cropped image would only represent less than 1000 m^2 of surface area. In many cases, this would only cover a small part of a landslide, preventing the network model from effectively processing the landslide's regional context information [39]. Meanwhile, using a resolution of 0.03 m and cropping to 1024 × 1024 size would result in a dataset with over 50,000 images, which would be a significant burden for training. Therefore, using a dataset with a resolution of 0.1 m could effectively address the two aforementioned issues.

2.2. Dataset Creation

There are three types of image data used as inputs for the semantic segmentation model, which can be summarized as follows: optical semantic data, topographic semantic data, and label data. The optical semantic data are represented by orthophoto, while the label data are obtained by landslide experts who observe the orthophoto and annotate it using ArcGIS software. Here, it is worth mentioning the selection of topographic semantic data. We ultimately chose relative elevation, slope, and hillshade as the three types of data to collectively constitute the topographic semantic data used for input (Figure 4).

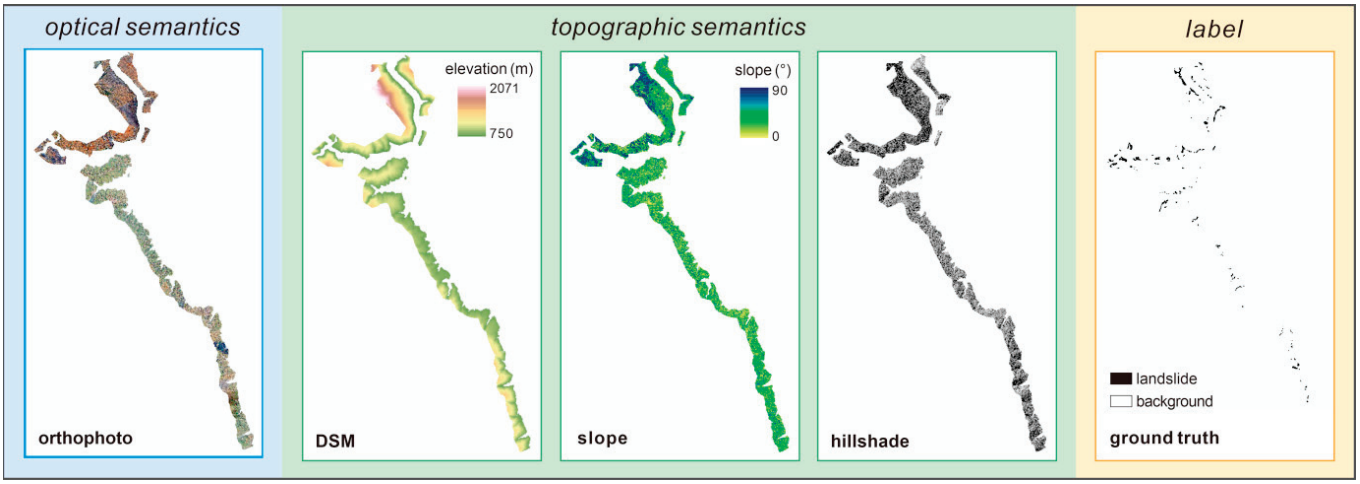


Figure 4. Input image data used for the semantic segmentation model. The orthophoto is composed of three channels including red (R), green (G), and blue (B) captured by UAV aerial camera, and the relative elevation, slope, hillshade and ground truth are composed of a single channel.

Relative elevation measures the degree of elevation difference within a region, numerically represented by the absolute elevation difference between the highest and lowest points in the area. The relative elevation metric directly affects the gravitational potential energy level. The greater the maximum value of relative elevation within an area, the higher the potential energy at that point, indicating an increase in the driving force of the landslide [40,41]. Although the concept of relative elevation is intended to be used, in practice, when composing topographic semantic data, we directly use elevation (i.e., the DSM itself) as input, because neural networks inherently perform normalization operations when processing data inputs, and the processed elevation becomes relative elevation.

Slope represents the angle between the vertical height of a point and its horizontal projection on a slope, intuitively reflecting the degree of slope inclination, which is also importantly related to landslides. An increase in the slope means that the component of gravity projected onto the potential sliding surface of the landslide also increases [42]. According to statistical data, most landslides have a slope between 10° and 30° [43].

By simulating the position and angle of the sun's light source, hillshade calculates the illumination of each pixel to reflect the undulations and terrain of the landscape. Hillshade

does not have a strong correlation with landslides on a geological level, but it has unique advantages when analyzing landslides using UAV datasets: it can reduce the impact of illumination differences in UAV data at different times [44]. In fact, the DSM exported by SfM can also mitigate the effects of illumination differences, but compared to the RGB image, the DSM loses a lot of shape and texture information. However, the hillshade calculated from the DSM can supplement a significant amount of image shape and texture, making it an ideal source of topographic semantic data for landslides. In this study, both slope and hillshade were directly exported from the DSM using ArcGIS software version 10.8. The solar altitude angle and the orientation of hillshade in this manuscript are 45° and 90°, respectively.

Furthermore, it is important to note that in subsequent sections of this paper, the term "topographic image" is used as a generic term for the DSM, hillshade, and slope, while the term "optical image" refers to the orthophoto. Both the terms precisely convey the core content of these two types of data, and it is helpful for understanding the data usage in this study.

To ensure that the model training could run smoothly under limited hardware conditions, the orthophoto, DSM, slope, hillshade, and label data were then cropped to a size of 1024 × 1024 and the cropped image set was filtered. The filtered images met at least one of the following conditions: (1) the proportion of pixels with null values exceeded 90%; (2) there were no pixels labeled as landslides. After the above operations, a dataset composed of 633 images of 1024 × 1024 size was obtained. The dataset was randomly divided into a training set and a validation set in an 8:2 ratio, which were used for model training and validation, respectively.

The entire dataset creation process for this section is shown in Figure 5, which also includes the mathematical expression of the filtering conditions for the image set.

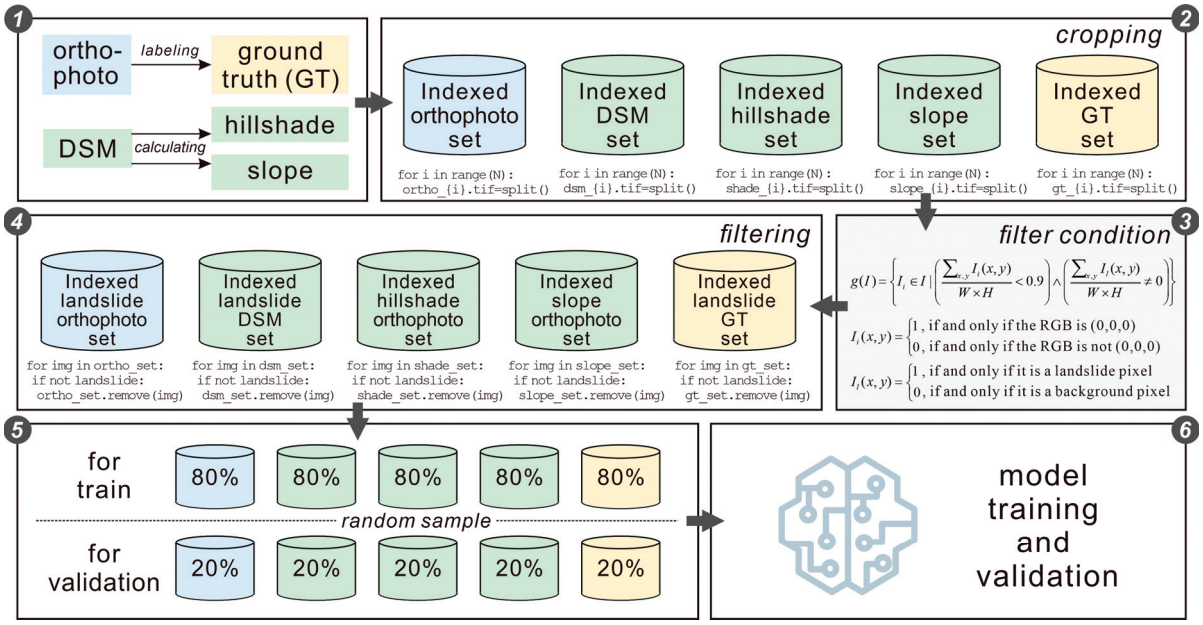


Figure 5. Workflow of dataset creation from step 1 to step 6. In step 1, the GT was labeled by manually observing the landslide areas in the orthophoto, and the hillshade and slope were exported from DSM. In step 2, certain image cropping was performed (pseudo code has been presented), and the large orthophoto was converted into an indexed set consisting of a series of small-scale images. In step 3, the filter condition is presented, and I is the small image after cropping, (x, y) is the coordinate of a pixel, W and H are the weight and height of the image before cropping, respectively, and $g(I)$ is the dataset after filtering. In step 4, the filtering was applied according to the filter condition (pseudo code has been presented). In step 5, each indexed set was divided into two parts for the training and validation procedures, respectively. In step 6, the final dataset was fed into the model for training and validation.

2.3. Dataset Augmentation

In order to expand the dataset, avoid model overfitting, and enhance the generalizability of the model, data augmentation (DA) methods were used to improve the training set [45]. Specifically, four types of DA operations were applied: stretching, flipping, Gaussian blur, and rotation (Figure 6). We applied these four transformations using an on-the-fly data augmentation approach [46]. During the training process, these DA operations were applied randomly before the dataset was loaded into memory. Theoretically, this means that the training set for each epoch is different during the training process, increasing the diversity of the data and helping the model generalize better to unseen data.

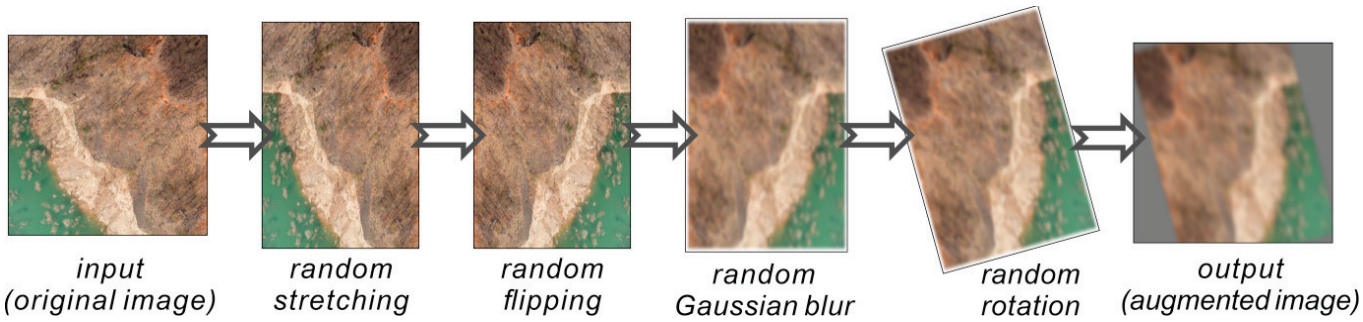


Figure 6. Dataset augmentation procedure in the dataset loading process of each epoch. The original will go through a sequential process consisting of stretching, flipping, Gaussian blur and rotation procedures. All procedures were applied randomly and were designed to increase the randomness of the final augmented image.

3. Methods

3.1. DeepLabV3+

3.1.1. Model Architecture

DeepLabV3+ is a deep learning semantic segmentation model developed by Google, which has been widely applied in fields such as medical image segmentation [47], remote sensing image analysis [48], and autonomous driving [49]. Compared to previous models in the DeepLab series, DeepLabV3+ introduces a decoder module to refine segmentation results, especially along the boundaries of target objects, making it also a typical semantic segmentation model with an encoder-decoder architecture. DeepLabV3+ also upsamples the output feature maps from the encoder through the decoder module and fuses them with low-level features, thereby restoring detailed information about the targets and improving the accuracy of segmentation boundaries. A simplified diagram of the overall architecture of DeepLabV3+ is shown in Figure 7.

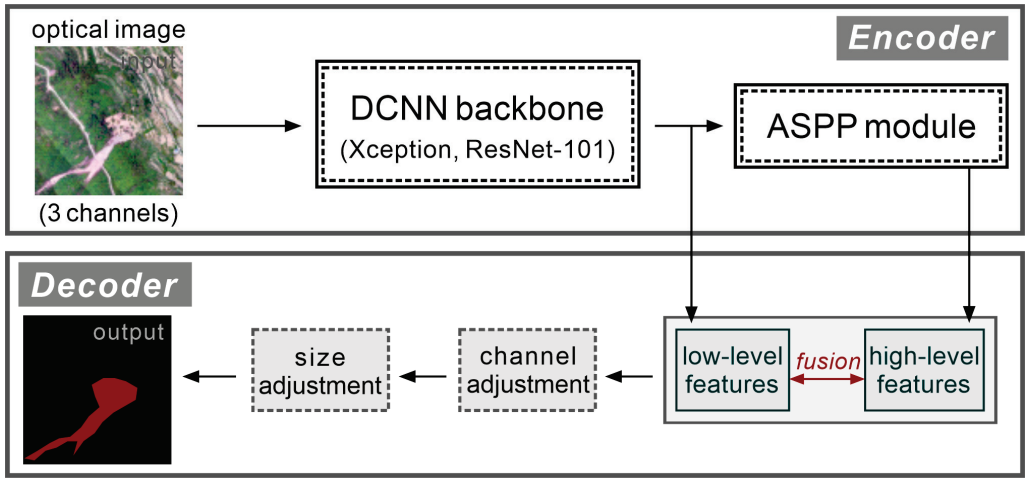


Figure 7. The basic structure of the DeepLabV3+ model, showing a typical encoder-decoder architecture.

3.1.2. The Dilated Convolution & ASPP Optimizations

The two most important optimizations in the DeepLabV3+ encoder are dilated convolutions and Atrous Spatial Pyramid Pooling (ASPP), which indirectly contribute to DeepLabV3+'s excellent performance in multi-scale object recognition. By using dilated convolution operations instead of pooling operations, it is possible to obtain a receptive field larger than the size of the convolutional kernel without reducing the size of the feature map, while keeping the number of parameters manageable. For a standard convolution process, it can be described as:

$$y[i,j] = \sum_{m=-M/2}^{M/2} \sum_{n=-N/2}^{N/2} x[i+m, j+n] \cdot w[m,n] + b, \quad (1)$$

However, in dilated convolution, the distance between adjacent pixels in the image corresponding to the convolutional kernel is not constantly 1, which means that:

$$y[i,j] = \sum_{m=-M/2}^{M/2} \sum_{n=-N/2}^{N/2} x[i+r \cdot m, j+r \cdot n] \cdot w[m,n] + b, \quad (2)$$

where y is the element of the output feature map, x is the element of the input feature map, w is the weight of the convolution kernel, b is the bias, m and n are the indexes of the convolution kernel, and r is the dilated ratio, which is an integer greater than or equal to 1.

Dilated convolution can effectively improve the ability of the model to capture the features of large targets, such as landslides. As shown in Figure 8, when the dilation rate of the 3 × 3 convolution kernel is 4, the receptive field of the convolution operation is expanded from a space of 9 pixels to a space of 81 pixels, which already contains almost the whole landslide area.

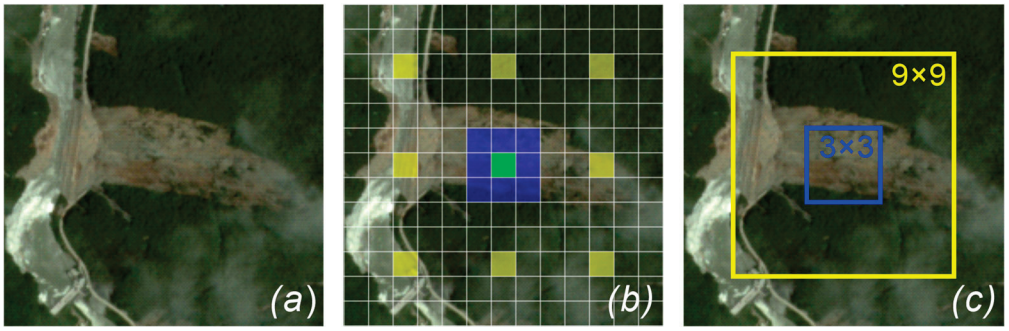


Figure 8. The dilated convolutions with a dilation rate of 4 bring an expanded receptive field to a 3 × 3 convolutional kernel: (**a**) an orthophoto including an entire landslide; (**b**) feature pixels corresponding to dilated and standard convolutions (yellow: feature pixels for dilated convolution; blue: feature pixels for standard convolution; green: feature pixels for both dilated and standard convolution); (**c**) schematic of the receptive field area.

Relying solely on dilated convolution can effectively expand the receptive field, but it does not achieve unified feature extraction for multi-scale objects. To address the issue of multi-scale feature extraction, Spatial Pyramid Pooling (SPP) has provided an excellent practical approach—establishing multiple parallel branches and using pooling operations with different kernel sizes in each branch to extract and ultimately integrate multi-scale features [50]. On the other hand, as mentioned above, dilated convolutions can be seen as a non-downsampling alternative to pooling operations. Naturally, the idea arises to use dilated convolution operations to replace the pooling operations in some branches of SPP, giving rise to Atrous Spatial Pyramid Pooling. Since dilated convolutions do not change the size of the feature map, this substitution allows the output results of each branch to be directly concatenated in the channel dimension (except for the pooling branches, which require upsampling), resulting in a multi-scale feature map that fully accounts for the features of objects at different scales. Figure 9 shows the basic structure of the ASPP module.

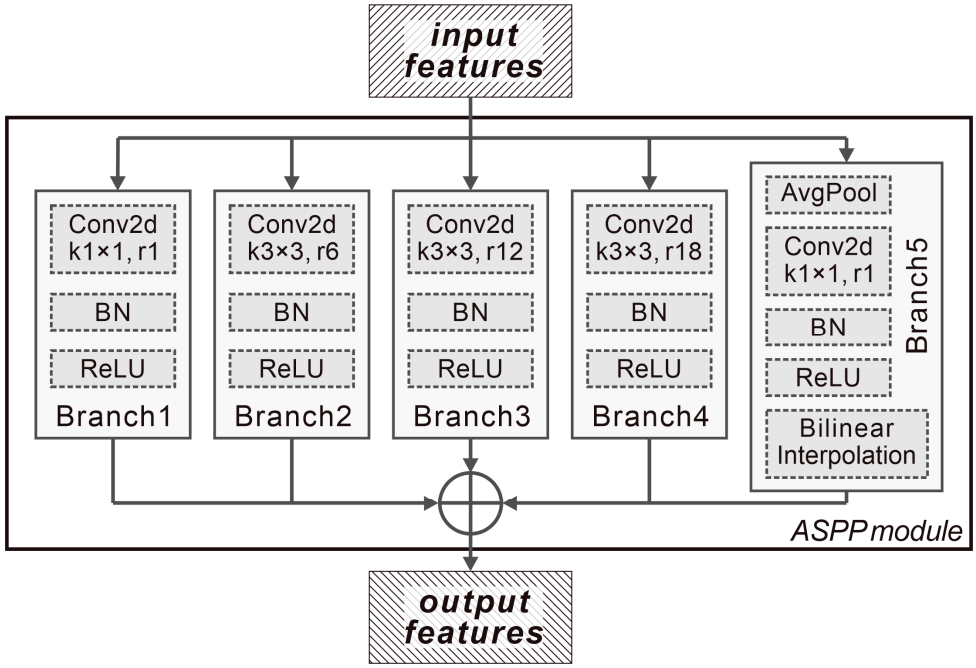


Figure 9. The basic structure of the ASPP module.

DeepLabV3+ has shown good performance on several open-source datasets, but these datasets are based on RGB optical images obtained from common imaging devices.

Therefore, the topographic information in the UAV datasets would be lost by using the general DeepLabV3+. For this reason, in order to achieve a landslide semantic segmentation model with good generalization performance, we re-designed a new DeepLabV3+-based semantic segmentation model for landslide recognition that includes topographic information.

3.2. DeepLab4LS: A Dual-Encoder DeepLab Model for Landslide

3.2.1. Basic Design of Model

In order to design a landslide recognition model that fuses both optical and topographic semantics, it is first necessary to consider the fusion strategy. In deep learning tasks for computer vision, common data fusion methods can be categorized into three types: simple fusion, channel fusion, and dual-encoder fusion (Figure 10). Simple fusion is considered the most basic form of data fusion, where two sets of data are directly added or multiplied together. This method necessitates high consistency between the two datasets and is typically used for fusing intermediate features within the network [51–53]. Channel fusion concatenates optical and topographic data along the channel dimension before its input into the model, thus better preserving the original characteristics of the data [54–56]. Dual-encoder fusion (also known as dual-backbone fusion) goes a step further by feeding each data set into an encoder (which is primarily composed of a backbone) and the features outputted by the encoders are then fused at the model level (usually through direct channel fusion) [57–59]. Compared to the other methods, dual-encoder fusion is the most complex, but it allows the data to undergo several layers of neural network processing both before and after fusion, thereby enabling a true deep fusion of data.

Therefore, to achieve better performance, this study employs a dual-encoder fusion approach for the model design. Overall, the network receives two types of inputs: orthophotos as optical features, and DSM, slope, and hillshade images as topographic features. Initially, two separate encoders are used to extract preliminary optical and topographic semantic features. Subsequently, these extracted semantic features are combined through channel fusion, followed by channel adjustment, size adjustment, and SoftMax processing to produce the final segmentation results. Throughout this process, the subsampling rates

of the backbone within the encoders are appropriately adjusted to unify the subsampling multiples of the dual encoders.

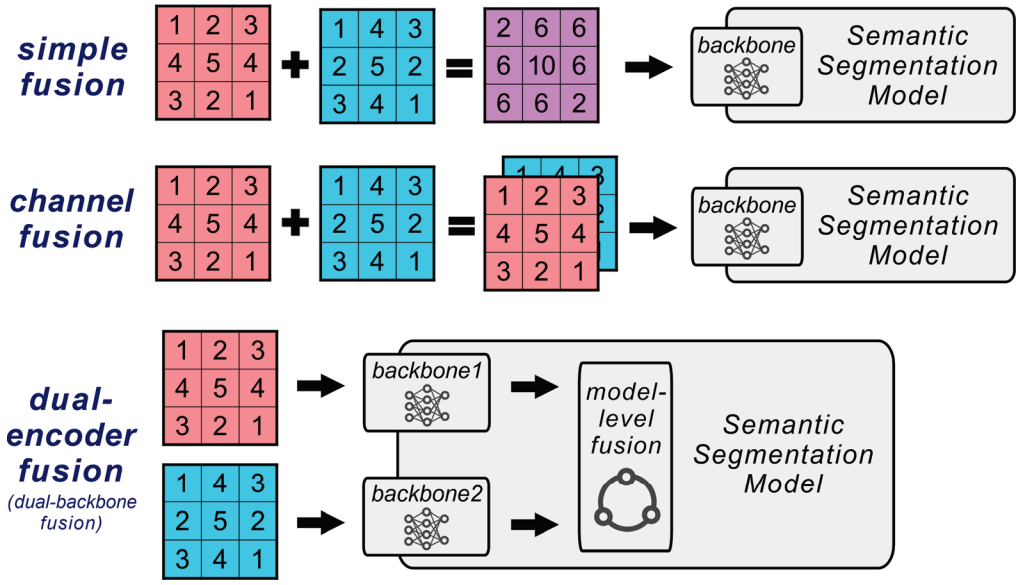


Figure 10. The sketch map of common data fusion methods: simple fusion, channel fusion, and dual-encoder fusion.

3.2.2. Encoders Design

Considering the increase in the number of parameters introduced by the new encoders, the optical feature extraction backbone adopts the lightweight MobileNetV2 [60]. MobileNetV2 achieves model miniaturization and efficient inference speed through its unique inverted residual structure combined with depthwise separable convolutions. In numerous experiments [61–63], MobileNetV2, as a backbone, has been able to achieve results comparable to or even surpassing those of ResNet50 [64], with the former having only 1/7 to 1/8 of the latter's parameters. Overall, MobileNetV2 can be regarded as an excellent lightweight backbone in terms of both computational accuracy and parameter volume.

With regard to the topographic feature extraction backbone, no single network emerges as a clear standout. In order to identify the most suitable network for this purpose, this study tested five different networks for their ability to extract topographic semantics, including VGG-16 [65], ResNet-18, ResNet-50, ShuffleNetV2 [66], and MobileNetV2. Among these, ShuffleNetV2 is a lightweight network structure similar to MobileNetV2. In ShuffleNetV2, the introduction of channel shuffle operations and pointwise group convolution techniques significantly reduces the model's parameter count while ensuring high accuracy. Regardless of which topographic feature extraction backbone is used, from the perspective of controlling the overall model's parameter volume, none employs the deep and shallow feature fusion method found in DeepLabV3+. Instead, they directly process based on the final output.

3.2.3. Model Architecture

With this understanding, we referenced the architecture of DeepLabV3+ and redesigned and implemented a recognition semantic segmentation model for landslide recognition—DeepLab for Landslide (DeepLab4LS) (Figure 11), composed of four modules: Encoder1, Encoder2, Mixer, and Decoder. The detailed structural design of the four modules is as follows:

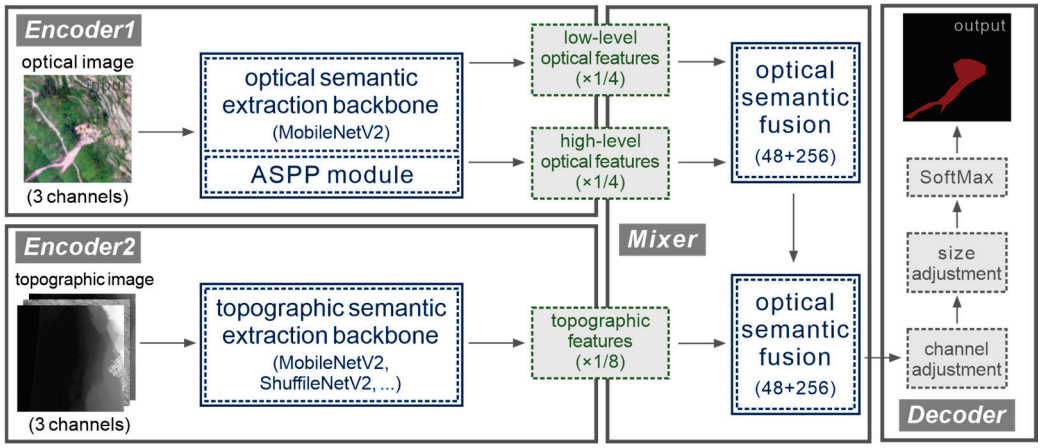


Figure 11. The basic structure of the DeepLab for Landslide (DeepLab4LS) model, composed of Encoder1, Encoder2, Mixer and Decoder.

(1) Encoder1

Encoder1 serves as the optical feature extraction module, comprising the optical feature extraction backbone and its associated components. Encoder1 takes RGB images (orthophotos) as input and normalizes them to a range of [0, 1] before entering the network. The optical feature extraction backbone is based on the MobileNetV2 model, with the original model's downsampling factor adjusted to 8× to prevent data distortion caused by excessive downsampling. The final output of Encoder1 is divided into two parts: 48-channel 8× downsampled shallow optical semantic features and 256-channel 8× downsampled deep optical semantic features.

(2) Encoder2

Encoder2 acts as the topographic feature extraction module, comprising the topographic feature extraction backbone and its associated components. Encoder2 accepts a 3-channel image composed of stacked single-channel images of DSM, slope, and hillshade as input, and normalizes it to a range of [0, 1] before entering the network. In order to align with the deep optical semantic features outputted by Encoder1, the downsampling factor of Encoder2 is also set to 8x, and the final channel count of topographic semantic features is fixed at 256. This allows the shallow optical semantic features, deep optical semantic features, and topographic features to be fused at a ratio of 48:256:256.

(3) Mixer

The Mixer serves as the multimodal feature fusion module. Within the Mixer, the task is to further achieve cross-semantic fusion of optical and topographic semantics, building upon the foundation of fusing shallow and deep optical semantic features. Before fusion, deep optical semantic features and topographic semantic features are upsampled by a factor of 2 using bilinear interpolation to return to a 4× downsampled size. The three sets of features are then channel-stacked and immediately processed through two 3×3 convolution layers. The first convolution layer transforms the 256 + 48 + 256 channel features into 256 channels of fused semantics, while the second convolution layer further refines the fused semantics without changing the number of channels. As a result, the Mixer outputs a 256-channel feature map downsampled by 4×, representing the outcome of deep multimodal feature fusion.

(4) Decoder

The Decoder is the conversion module for segmentation results. The main function of the Decoder is to convert the feature map output from the Mixer into the final segmentation results. Therefore, its network structure is relatively simple, consisting of a 1×1 convolution layer, a bilinear interpolation module, and a final SoftMax module. The 1×1

convolution layer is used for channel adjustment (outputting 2 channels, representing the landslide foreground and background), the bilinear interpolation adjusts the image size to match the input image size, and the subsequent SoftMax transformation yields pixel classification probabilities. Finally, the category with the highest probability is selected as the pixel class to produce the final segmentation results.

It is important to emphasize that in the aforementioned Encoder1 and Encoder2 modules, adjustments to the downsampling multiples of each model are made by modifying the stride of the convolutional or pooling layers accordingly. Another modification to these models is the removal of the classifiers, so that they exist only as a backbone for a semantic segmentation model, rather than as complete classification models.

3.2.4. Improved Loss Function

In this study, another significant endeavor is the formulation of a loss function tailored for landslide recognition, which is then applied to the DeepLab4LS model. Given the relatively low proportion of landslide pixels in the image, a novel loss function, positive enhanced loss (*PE loss*), is introduced. This loss function is designed to enhance the model's ability to comprehend positive samples. The *PE loss* is defined by the following equation:

$$PE\ loss = 1 - \frac{1}{|C|} \frac{\sum\limits_{pixels} y_{\text{true}} y_{\text{pred}}}{\sum\limits_{pixels} \left(y_{\text{pred}} + y_{\text{true}} - 2 y_{\text{true}} y_{\text{pred}} \right)}, \tag{3}$$

Here, y_{true} is equal to 1 when a pixel is a positive sample and 0 otherwise, while y_{pred} is equal to 1 when a pixel is predicted as a positive sample and 0 otherwise. The value of $|C|$ is the total number of classes, which is consistently equal to 2 in this study.

During the training process of the DeepLab4LS model, the loss function is composed of both Positive Enhanced loss and Cross Entropy loss (*CE loss*), which is:

$$loss = PE\ Loss + CE\ Loss, \tag{4}$$

CE loss is a commonly used loss function in deep learning. The official implementation of DeepLabV3+ employs *CE loss* as its loss function [67]. The incorporation of the PE loss component into the loss function serves to reinforce the gradient stability and to enhance the model's probability estimation capabilities.

3.3. Training Strategy

To facilitate model convergence and maintain the stability of learning low-level features, this study adopts a strategy of using pre-trained weights combined with a freeze-training approach. Given that the DeepLab4LS model's backbones are based on widely-used feature extraction networks that have been slightly adjusted for downsampling factors, it is possible to apply the pre-trained weights from these networks, which have been trained on various open-source datasets, as the initial weights for DeepLab4LS's backbone components. The initial weights, w_i, of the non-backbone parts of DeepLab4LS are initialized to follow a normal distribution with a mean of zero and a standard deviation of 0.02.

$$w_i \sim N\left(0.0, 0.02^2\right), \tag{5}$$

The model training was conducted using an NVIDIA GeForce RTX 3060 graphics card with 12G of dedicated VRAM. The training phase was set to a total of 500 epochs, with the first 50 epochs designated for freeze training. During the freeze phase, the batch size was set to 8, and for the unfreeze phase, it was reduced to 4. This approach ensured that the total number of training steps exceeded 79,000, guaranteeing full convergence of the model. In the initial freeze training phase, the weights of the backbone components (Encoder1 and Encoder2) were fixed, allowing for rapid adjustment of the Mixer and Decoder weights, which quickly elevated the model to a higher level of accuracy. In the subsequent unfreeze

phase, the weights of the entire model, including the backbone networks, were fine-tuned, enabling the model to gradually develop the capability for deep extraction and inference of landslide semantics.

Furthermore, to minimize VRAM consumption, the training process employs PyTorch's mixed-precision training strategy. This strategy can boost training efficiency by integrating half-precision (float16) with single-precision (float32) floating-point numbers, thus striking a balance between performance and memory usage. Additionally, key hyperparameters that influence the training dynamics, including optimizer and learning rate, are concisely detailed in Table 1.

Table 1. The training parameters set for DeepLab4LS.

Environment and Parameters	Name and Value
Development Environment	Python 3.7.13/PyTorch 1.10.1/NVIDIA CUDA 11.3
Epoch	50 (freeze) + 450 (unfreeze)
Batch Size	8 (freeze)/4 (unfreeze)
Optimizer	Stochastic gradient descent
Momentum	0.9
Weight Decay	10^{-4}
Learning Rate	7×10^{-3} (beginning)/7×10^{-5} (minimum)
Learning Rate Decay	Cosine annealing

3.4. Evaluation Metrics

In this study, mIoU (mean Intersection over Union), mPA (mean Pixel Accuracy) and Accuracy were introduced to evaluate the accuracy of the semantic segmentation model in landslide recognition tasks. mIoU is a crucial metric for almost all semantic segmentation models and indicates the percentage of intersection over the union of the predicted and ground truth regions, providing a quick assessment of the model's performance across all classes. mPA is derived by averaging the percentage of correctly classified pixels within each class, placing more emphasis on the model's performance in each individual class. Accuracy, which is more intuitive than the first two, represents the proportion of correctly predicted pixels out of the total pixels, reflecting the model's performance across the entire dataset. Specifically, the calculation methods for these three evaluation metrics are as follows:

$$\text{mIoU} = \frac{1}{N}\sum_{i=1}^{N} \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i + \text{FN}_i}, \tag{6}$$

$$\text{mPA} = \frac{1}{N}\sum_{i=1}^{N} \frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i}, \tag{7}$$

$$\text{Accuracy} = \frac{\sum_{i=1}^{N} \text{TP}_i + \text{TN}_i}{\sum_{i=1}^{N} \text{TP}_i + \text{TN}_i + \text{FP}_i + \text{FN}_i}, \tag{8}$$

where N is the total number of classes, and TP_i, FP_i, TN_i, FN_i are the true positive, false positive, true negative, and false negative counts for the i-th class, respectively.

4. Results

4.1. Segmentation Results and Comparison Experiments

The segmentation results of DeepLab4LS using different topographic semantics to extract the backbone are shown in Table 2. It is evident that no single topographic semantic extraction backbone excels in all three evaluation metrics simultaneously. However, in the overall evaluation, MobileNetV2 emerges as the superior topographic feature extraction backbone network. It has the highest mIoU and the second highest mPA and precision, demonstrating its robust ability to extract landslide semantics from topographic features. Somewhat surprisingly, VGG-16 achieved the highest mPA of the five topographic semantic extraction backbones, but scored the lowest in both mIoU and accuracy. This suggests

that the model may have a tendency to predict pixels as landslides, or it may not have effectively learned the boundary features of landslides. Overall, considering both number of parameters and performance, MobileNetV2 stands out as a relatively optimal backbone for topographic semantic extraction.

Table 2. Segmentation results of DeepLab4LS under different topographic semantic extraction backbones.

Optical Semantic Extraction Backbone	Topographic Semantic Extraction Backbone	mIoU (%)	mPA (%)	Accuracy (%)
MobileNetV2	MobileNetV2	**76.0**	85.3	92.3
	VGG-16	72.6	**86.3**	90.3
	ShuffleNetV2	75.5	83.7	**92.4**
	ResNet-18	74.1	84.1	91.5
	ResNet-50	74.8	83.7	92.0

Further, the DeepLab4LS model was compared with five other popular semantic segmentation models, and the results are shown in Table 3. It is quite apparent that the DeepLab4LS model has an absolute advantage in all evaluation metrics, with its mIoU surpassing the second place by 5.5 percentage points, indicating the significant gain brought by the integration of topographic features in landslide recognition. Among the other five models, DeepLabV3+ and PSPNet exhibited similar performance and were superior to the other three semantic segmentation models. Notably, U-Net, which excels in medical imaging datasets, performed moderately on the landslide drone dataset. In fact, models like DeepLabV3+ and PSPNet [68] have adopted a degree of multi-scale optimization to enhance the model's multi-scale target perception capabilities; SegFormer [69] also employed a Transformer architecture for more effective regional feature modeling. In contrast, U-Net has a simpler architecture and lacks sufficient multi-scale target extraction capabilities, which may be the main reason limiting its landslide recognition ability.

Table 3. Comparison experiments between DeepLab4LS and other popular semantic segmentation models.

Model	mIoU (%)	mPA (%)	Accuracy (%)
DeepLab4LS (with MobileNetV2 as the topographic backbone)	**76.0**	**85.3**	**92.3**
DeepLabV3+	70.5	80.1	90.4
PSPNet	70.5	81.2	90.1
SegFormer	70.1	77.8	90.8
U-Net	67.6	77.5	89.3
HRNet	69.6	80.0	89.8

In order to visually analyze the performance enhancement in landslide recognition exhibited by DeepLab4LS compared to the DeepLabV3+ model, we selected a subset of images from the validation set for comparative analysis, as shown in Figure 12. It can be observed that DeepLab4LS shows a significant enhancement in landslide recognition performance over the DeepLabV3+ model. For landslides with distinct boundary information, both DeepLab4LS and DeepLabV3+ can roughly identify the extent of the landslides (as in rows 3, 5, and 8 of Figure 12); however, for those landslides with less distinct boundaries (as in rows 6 and 7 of Figure 12), DeepLabV3+ often struggles to produce reliable results. It is also noteworthy that even for landslides with distinct boundaries, the edges recognized by DeepLabV3+ tend to be coarse. In contrast, DeepLab4LS is able to delineate the true boundaries of the landslides with greater precision.

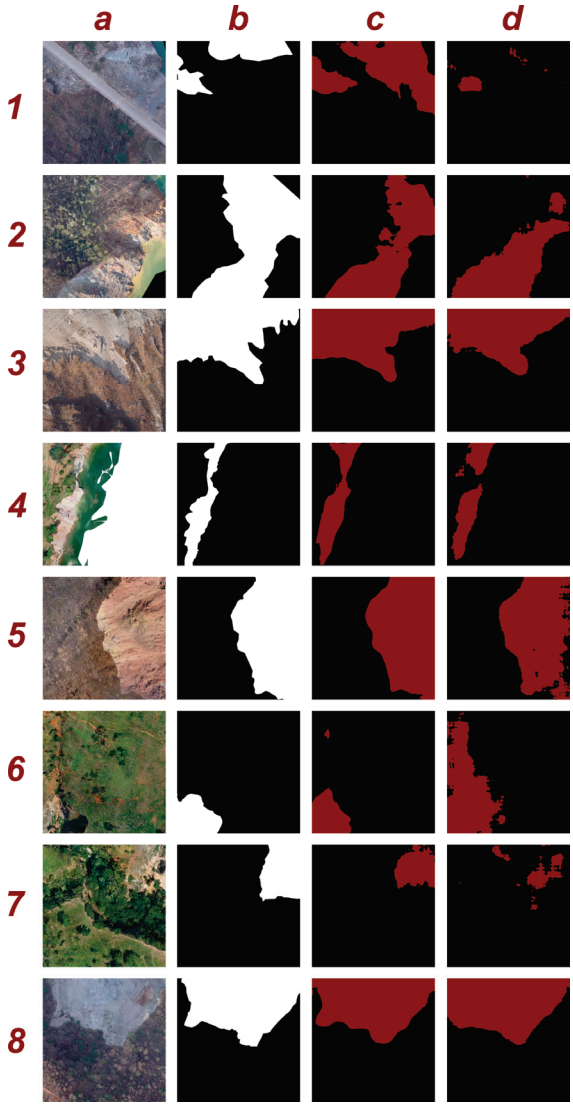


Figure 12. The results of the eight typical segmentation results numbered as 1–8: (**a**) the original UAV orthophoto; (**b**) the ground truth determined by landslide experts; (**c**,**d**) the segmentation results obtained by DeepLab4LS model and DeepLabV3+ model, respectively.

4.2. Ablation Experiments

In this study, we designed the DeepLab4LS model, which consists of Encoder1, Encoder2, Mixer and Decoder modules, and introduced the PE loss function to enhance the model's ability to understand positive samples. Through a series of ablation experiments, we aim to demonstrate the contributions of these modules or the loss function. Specifically, we aim to demonstrate:

(1) The contribution of the new combination of loss functions.
(2) The contribution of the Mixer module.
(3) The contribution of the dual-encoder architecture.

The design and results of the ablation experiments are presented in Table 4, revealing significant insights. The loss function in this study is composed of the CE and PE terms, and the removal of either significantly diminishes the performance of the model. Both the Mixer and the dual-encoder play a crucial role in the model; removing the Mixer and Encoder2 results in a decrease of 1.6% and 1.0% in mIoU, respectively. The contribution of Encoder1 is even more pronounced; without it, the model's mIoU drops by 12.1%, which is clearly unacceptable. Nevertheless, the ablation experiments have successfully demonstrated the

vital contributions of the combined loss functions, the Mixer module, and the dual-encoder architecture to the model.

Table 4. Ablation experiments based on the DeepLab4LS model, each experiment differing in module composition or loss function.

Base Model	Encoder1	Encoder2	Mixer	Loss Functions	mIoU (%)	mPA (%)	Accuracy (%)
DeepLab4LS (with MobileNetV2 as the topographic backbone)	✔	✔	✔	CE [2] + PE [3]	76.0	85.3	92.3
	✔	✔	✔	CE	72.2	81.2	91.1
	✔	✔	✔	PE	75.6	85.0	92.2
	✔	✔		CE + PE	74.4	81.7	92.2
		✔		CE + PE	63.9	76.4	83.9
	✔ [1]			CE + PE	75.0	84.0	92.1

[1] The DeepLab4LS model degrades to DeepLabV3+ model in this scenario (but with different loss functions). [2] CE: Cross Entropy (loss). [3] PE: Positive Enhanced (loss).

5. Discussion

5.1. Topographic Features in Landslide Segmentation

One of the major enhancements that DeepLab4LS offers over DeepLabV3+ is the incorporation of topographic images as part of the input. As indicated in Figure 12, the introduced topographic information aids in bolstering the model's comprehension of landslide semantics, particularly when landslide boundaries are indistinct. This outcome is unsurprising; as a geological entity, analyzing the topographic characteristics of landslides is crucial for their effective differentiation. Indeed, many features of landslides, such as cracks, fissures, accumulation bodies, and sliding surfaces, are more pronounced in topographic images [70,71], allowing the model to learn these features easily from the topographic images.

However, despite the significant assistance that topographic features in images provide for landslide recognition, ablation experiments (Table 4) indicate that relying solely on topographic images for recognition yields less accurate results. This phenomenon could be related to various factors, including digital imaging, neural networks, and the inherent characteristics of landslides themselves. Here, we propose a possible explanation: features in optical images are "easier" for neural networks to learn than those in topographic images. Therefore, when both optical and topographic images are used, the model can initially obtain some "hints" of landslide features from the optical images, which then synergistically promotes a deeper learning of landslide features from both optical and topographic images. In contrast, when only topographic images are used, the model is immediately confronted with challenging landslide features to learn, resulting in suboptimal training results.

5.2. Mixer of DeepLab4LS Model

The ablation experiments (Table 4) have confirmed the significant role of the Mixer in the model. It is noteworthy that when employing the dual-encoder structure without the Mixer, the mIoU is 74.4%, which is even lower than the 75.0% obtained when using only Encoder1. This indicates that merely introducing Encoder2 does not aid in a better understanding of the landslide semantics, but rather may somewhat impede the model's comprehension of the optical landslide semantics.

To gain a better understanding of the Mixer's function, we visualized the feature maps at five key points in DeepLab4LS by averaging the values across channels, as shown in Figure 13. It is readily observed that in the feature maps F1 and F2 output by Encoder1, the extent of landslide and background can be roughly discerned. However, in the feature map F3 output by Encoder2, the extent information is not as clear. This actually corroborates our conjecture in Section 5.1. that the topographic features of landslides are more challenging to learn. Once the three feature maps are stacked into F4, the network possesses both optical and topographic semantics of the landslide, but the features within the landslide extent remain somewhat disordered. Feature map F5 is the output of F4 after Mixer processing.

After this, feature map F5 is considerably smoother, with features within the landslide extent exhibiting high consistency. Finally, the Decoder captures such a feature distribution and outputs reliable segmentation results. In summary, the Mixer effectively facilitates the deep integration of optical and topographic semantics in the model.

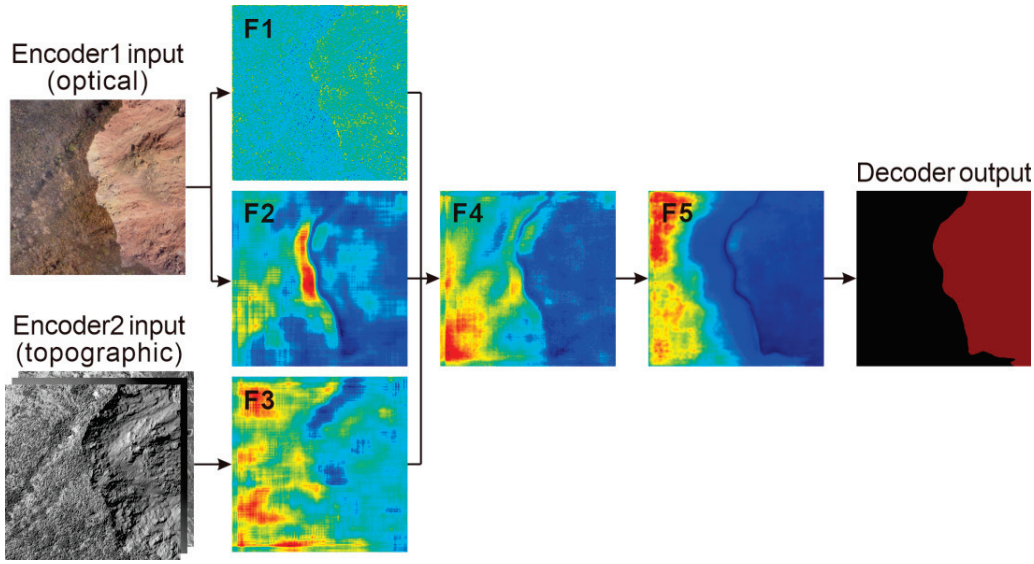


Figure 13. Visualization of feature maps for key nodes in DeepLab4LS. F1: low-level optical feature map of Encoder1 output; F2: high-level optical feature map of Encoder1 output; F3: topographic feature map of Encoder2 output; F4: concatenated feature map before fusion in Mixer; F5: fused feature map of Mixer output.

5.3. Loss Function in Landslide Segmentation

In this study, we attempted to enhance the model's comprehension of positive samples by introducing a novel loss function—PE loss. The results of the ablation experiments (Table 4) indicate that PE loss effectively achieved our goal. Although the standard loss function in this study was designed as a combination of CE loss and PE loss, the ablation results show that even when using PE loss alone, there is a significant performance enhancement over using CE loss alone. We can try to explain this phenomenon based on the formula of PE loss itself. Observing the fractional term of PE loss, the numerator represents the correctly predicted positive samples, while the denominator accounts for all incorrectly predicted samples. Compared to Soft IoU loss (which will be discussed in detail later), the denominator of PE loss lacks the correctly predicted positive samples, which means that PE loss will easily attain a small value when a large number of positive samples are successfully predicted and few positive and negative samples are mispredicted, thereby amplifying the penalty in the opposite scenario. In essence, through PE loss, we aim to instruct the model to focus its efforts on the current positive sample.

In fact, the design of PE loss in this study was inspired by another similar loss function—Soft IoU loss (*SI loss*) [72]. Compared to *PE loss*, *SI loss* includes correctly predicted positive samples in the denominator, making it "less aggressive" than *PE loss*. The classic form of *SI loss* is:

$$classic\ SI\ loss = 1 - \frac{1}{|C|} \frac{\sum\limits_{pixels} y_{\text{true}} y_{\text{pred}}}{\sum\limits_{pixels} \left(y_{\text{pred}} + y_{\text{true}} - y_{\text{true}} y_{\text{pred}} \right)}, \tag{9}$$

We can also logarithmize the fractional term, resulting in:

$$logarithmic\ SI\ loss = 1 - \log\left(\frac{1}{|C|}\frac{\sum\limits_{pixels} y_{\text{true}}y_{\text{pred}}}{\sum\limits_{pixels}\left(y_{\text{pred}} + y_{\text{true}} - y_{\text{true}}y_{\text{pred}}\right)}\right), \tag{10}$$

Clearly, logarithmic *SI loss* makes *SI loss* more "aggressive" in a more direct manner. Here, the meanings of the terms in Equations (9)–(11) are the same as those in Equation (3).

We applied the above two forms of *SI loss* to the training process, and the results are shown in Table 5. Additionally, we also tried to introduce *dice loss* [73], which is often used to address the issue of imbalance in samples, to see if it could be profitable for the training. The *dice loss* is defined as follows:

$$dice\ loss = 1 - \frac{1}{|C|}\frac{2\sum\limits_{pixels} y_{\text{true}}y_{\text{pred}}}{\sum\limits_{pixels}\left(y_{\text{pred}}^2 + y_{\text{true}}^2\right)} \tag{11}$$

Table 5. Segmentation results of DeepLab4LS under different loss functions.

Base Model	Loss Functions	mIoU (%)	mPA (%)	Precision (%)
DeepLab4LS (with MobileNetV2 as the topographic backbone)	CE [1]	72.7	81.5	91.3
	CE + PE [2]	**76.0**	**85.3**	**92.3**
	CE + dice	72.9	81.5	91.5
	CE + classic SI [3]	74.0	84.3	91.4
	CE + logarithmic SI [3]	72.2	82.5	90.8

[1] CE: Cross Entropy (loss). [2] PE: Positive Enhanced (loss). [3] SI: Soft IoU (loss).

Generally, the ranking of the loss functions in terms of performance gain for the model is as follows: *PE loss* (Equation (3)) > *classic SI loss* (Equation (9)) > *dice loss* (Equation (11)) > *logarithmic SI loss* (Equation (10)). *PE loss*, *dice loss*, and *classic SI loss* all improve the model's ability to recognize landslides to some extent. However, logarithmic *SI loss* resulted in a negative optimization, which may be due to its "excessive" enhancement of penalty, causing the model to easily fall into local optima. In summary, our practice has proven that *PE loss* can effectively enhance the performance of landslide recognition models.

6. Conclusions

In this study, we designed and implemented the DeepLab4LS model, a novel semantic segmentation model for landslide recognition based on DeepLabV3+. We introduced a dual-encoder design and established a model architecture composed of Encoder1, Encoder2, Mixer, and Decoder modules. Within this framework, the model receives both optical and topographic images as inputs and achieves deep integration within the model itself. This endows the model with the ability to infer both optical and topographic semantics of landslides, aligning more closely with the approach human experts use to recognize landslides. Another significant optimization in this study is the introduction of a new loss function term—Positive Enhanced loss (*PE loss*)—to enhance the model's comprehension of positive samples.

A series of comparison and ablation experiments were conducted on optical and topographic UAV datasets of the Baihetan reservoir. The results show that the DeepLab4LS model achieved an mIoU of 76.0%, outperforming DeepLabV3+'s mIoU of 70.5% and significantly surpassing other popular models such as SegFormer. These results are closely related to the introduction of the dual-encoder architecture and the Mixer module. Our experiments also found that optical and topographic images play a mutually reinforcing role, and that inputting only one type of image would lead to a decline in model performance. With only topographic images as input, the mIoU could drop by as much as 12.1%, which

is completely unacceptable. The introduction of *PE loss* is also crucial for enhancing model performance, increasing the mIoU by approximately 3.8 percentage points. In fact, PE loss allows the model to focus more on improving the recognition quality of current positive samples, which is evidently beneficial for landslide recognition tasks against complex backgrounds.

Author Contributions: All authors contributed to the manuscript and discussed the results. Z.-H.L. and H.-X.X. drafted the manuscript and were responsible for the data processing, analysis and interpretation of the results. A.-C.S. and Z.-H.N. proposed the ideas for the thesis, designed the structure and contributed to the final revision of the thesis. Y.-X.H. collected the on-site UAV data. H.-B.L. contributed to the UAV modeling and the SPOT analysis. N.J. revised the manuscript and contributed to the final revision of the thesis. All authors have read and agreed to the published version of the manuscript.

Funding: This study has been supported by the National Natural Science Foundation of China (U2240221), the Sichuan Youth Science and Technology Innovation Research Team Project (2020JDTD0006) and the Key Science and Technology Plan Project of PowerChina Huadong Engineering Corporation Limited (KY2021–ZD–03).

Data Availability Statement: The data presented in this study are available on request from the corresponding author due to confidentiality requirements.

Acknowledgments: The authors would like to thank the anonymous reviewers for their time and constructive comments on our article.

Conflicts of Interest: Author Zhi-Hai Li, An-Chi Shi and Zi-Hao Niu were employed by the company Power China Huadong Engineering Corporation Limited, Hangzhou 311122, China. 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. Hungr, O.; Leroueil, S.; Picarelli, L. The Varnes classification of landslide types, an update. *Landslides* **2014**, *11*, 167–194. [CrossRef]
2. Highland, L.M.; Bobrowsky, P. *The Landslide Handbook—A Guide to Understanding Landslides*; US Geological Survey: Reston, CA, USA, 2008.
3. Fang, K.; Dong, A.; Tang, H.; An, P.; Wang, Q.; Jia, S.; Zhang, B. Development of an easy-assembly and low-cost multismartphone photogrammetric monitoring system for rock slope hazards. *Int. J. Rock Mech. Min.* **2024**, *174*, 105655. [CrossRef]
4. Zhang, J.; Gurung, D.R.; Liu, R.; Murthy, M.S.R.; Su, F. Abe Barek landslide and landslide susceptibility assessment in Badakhshan Province, Afghanistan. *Landslides* **2015**, *12*, 597–609. [CrossRef]
5. Peng, M.; Ma, C.; Shen, D.; Yang, J.; Zhu, Y. Breaching and Flood Routing Simulation of the 2018 Two Baige Landslide Dams in Jinsha River. In *Dam Breach Modelling and Risk Disposal: Proceedings of the First International Conference on Embankment Dams (ICED 2020) 1*; Springer: Cham, Switzerland, 2020; pp. 371–373.
6. Yang, W.; Fang, J.; Jing, L.-Z. Landslide-lake outburst floods accelerate downstream slope slippage. *Earth Surf. Dyn. Discuss.* **2021**, *9*, 1251–1262. [CrossRef]
7. Li, H.; Li, X.; Ning, Y.; Jiang, S.; Zhou, J. Dynamical process of the Hongshiyan landslide induced by the 2014 Ludian earthquake and stability evaluation of the back scarp of the remnant slope. *Bull. Eng. Geol. Environ* **2019**, *78*, 2081–2092. [CrossRef]
8. Zhou, J.; Lu, P.; Yang, Y. Reservoir landslides and its hazard effects for the hydropower station: A case study. In *Advancing Culture of Living with Landslides: Volume 2 Advances in Landslide Science*; Springer: Cham, Switzerland, 2017; pp. 699–706.
9. Fell, R.; Hartford, D. *Landslide Risk Management. Landslide Risk Assessment*; Routledge: London, UK, 2018; pp. 51–109.
10. Pazzi, V.; Morelli, S.; Fanti, R. A review of the advantages and limitations of geophysical investigations in landslide studies. *Int. J. Geophys.* **2019**, *2019*, 2983087. [CrossRef]
11. Giordan, D.; Adams, M.S.; Aicardi, I.; Alicandro, M.; Allasia, P.; Baldo, M.; De Berardinis, P.; Dominici, D.; Godone, D.; Hobbs, P. The use of unmanned aerial vehicles (UAVs) for engineering geology applications. *Bull. Eng. Geol. Environ.* **2020**, *79*, 3437–3481. [CrossRef]
12. Rothmund, S.; Niethammer, U.; Malet, J.; Joswig, M. Landslide surface monitoring based on UAV-and ground-based images and terrestrial laser scanning: Accuracy analysis and morphological interpretation. *First Break* **2013**, *31*. [CrossRef]
13. Hu, S.; Qiu, H.; Pei, Y.; Cui, Y.; Xie, W.; Wang, X.; Yang, D.; Tu, X.; Zou, Q.; Cao, P. Digital terrain analysis of a landslide on the loess tableland using high-resolution topography data. *Landslides* **2019**, *16*, 617–632. [CrossRef]
14. Furukawa, F.; Laneng, L.A.; Ando, H.; Yoshimura, N.; Kaneko, M.; Morimoto, J. Comparison of RGB and multispectral unmanned aerial vehicle for monitoring vegetation coverage changes on a landslide area. *Drones* **2021**, *5*, 97. [CrossRef]

15. Bui, T.; Lee, P.; Lum, K.; Loh, C.; Tan, K. Deep learning for landslide recognition in satellite architecture. *IEEE Access* **2020**, *8*, 143665–143678. [CrossRef]
16. Li, Z.; Jiang, N.; Shi, A.; Zhao, L.; Xian, Z.; Luo, X.; Li, H.; Zhou, J. Reservoir landslide monitoring and mechanism analysis based on UAV photogrammetry and sub-pixel offset tracking: A case study of Wulipo landslide. *Front. Earth Sci.* **2024**, *11*, 1333815. [CrossRef]
17. Hölbling, D.; Eisank, C.; Albrecht, F.; Vecchiotti, F.; Friedl, B.; Weinke, E.; Kociu, A. Comparing Manual and Semi-Automated Landslide Mapping Based on Optical Satellite Images from Different Sensors. *Geosciences* **2017**, *7*, 37. [CrossRef]
18. Rosin, P.L.; Hervás, J. Remote sensing image thresholding methods for determining landslide activity. *Int. J. Remote Sens.* **2005**, *26*, 1075–1092. [CrossRef]
19. Yang, X.; Chen, L. Using multi-temporal remote sensor imagery to detect earthquake-triggered landslides. *Int. J. Appl. Earth Obs. Geoinf.* **2010**, *12*, 487–495. [CrossRef]
20. Eisank, C.; Smith, M.; Hillier, J. Assessment of multiresolution segmentation for delimiting drumlins in digital elevation models. *Geomorphology* **2014**, *214*, 452–464. [CrossRef]
21. Rau, J.; Jhan, J.; Rau, R. Semiautomatic object-oriented landslide recognition scheme from multisensor optical imagery and DEM. *IEEE Trans. Geosci. Remote Sensing* **2013**, *52*, 1336–1349. [CrossRef]
22. Liu, Q.; Wu, T.; Deng, Y.; Liu, Z. Intelligent identification of landslides in loess areas based on the improved YOLO algorithm: A case study of loess landslides in Baoji City. *J. Mt. Sci.* **2023**, *20*, 3343–3359. [CrossRef]
23. Cheng, L.; Li, J.; Duan, P.; Wang, M. A small attentional YOLO model for landslide detection from satellite remote sensing images. *Landslides* **2021**, *18*, 2751–2765. [CrossRef]
24. 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]
25. Soares, L.P.; Dias, H.C.; Grohmann, C.H. Landslide segmentation with U-Net: Evaluating different sampling methods and patch sizes. *arXiv* **2020**, arXiv:2007.06672.
26. Ding, P.; Zhang, Y.; Jia, P.; Chang, X. A comparison: Different DCNN models for intelligent object detection in remote sensing images. *Neural Process. Lett.* **2019**, *49*, 1369–1379. [CrossRef]
27. Singh, R.; Rani, R. Semantic segmentation using deep convolutional neural network: A review. In Proceedings of the International Conference on Innovative Computing & Communications (ICICC), Delhi, India, 21–23 February 2020.
28. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015; Proceedings, Part III 18; Springer: Cham, Switzerland, 2015; pp. 234–241.
29. 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.
30. Chen, L.; 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 (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
31. Yun, L.; Zhang, X.; Zheng, Y.; Wang, D.; Hua, L. Enhance the accuracy of landslide detection in UAV images using an improved Mask R-CNN Model: A case study of Sanming, China. *Sensors* **2023**, *23*, 4287. [CrossRef] [PubMed]
32. Fu, Y.; Li, W.; Fan, S.; Jiang, Y.; Bai, H. CAL-Net: Conditional Attention Lightweight Network for In-Orbit Landslide Detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 4408515. [CrossRef]
33. Ganerød, A.J.; Lindsay, E.; Fredin, O.; Myrvoll, T.; Nordal, S.; Rød, J.K. Globally vs. Locally Trained Machine Learning Models for Landslide Detection: A Case Study of a Glacial Landscape. *Remote Sens.* **2023**, *15*, 895. [CrossRef]
34. Pepe, M.; Fregonese, L.; Crocetto, N. Use of SfM-MVS approach to nadir and oblique images generated throught aerial cameras to build 2.5 D map and 3D models in urban areas. *Geocarto Int.* **2022**, *37*, 120–141. [CrossRef]
35. Dun, J.; Feng, W.; Yi, X.; Zhang, G.; Wu, M. Detection and mapping of active landslides before impoundment in the Baihetan Reservoir Area (China) based on the time-series InSAR method. *Remote Sens.* **2021**, *13*, 3213. [CrossRef]
36. He, C.; Hu, X.; Tannant, D.D.; Tan, F.; Zhang, Y.; Zhang, H. Response of a landslide to reservoir impoundment in model tests. *Eng. Geol.* **2018**, *247*, 84–93. [CrossRef]
37. Yi, X.; Feng, W.; Wu, M.; Ye, Z.; Fang, Y.; Wang, P.; Li, R.; Dun, J. The initial impoundment of the Baihetan reservoir region (China) exacerbated the deformation of the Wangjiashan landslide: Characteristics and mechanism. *Landslides* **2022**, *19*, 1897–1912. [CrossRef]
38. Cheng, Z.; Liu, S.; Fan, X.; Shi, A.; Yin, K. Deformation behavior and triggering mechanism of the Tuandigou landslide around the reservoir area of Baihetan hydropower station. *Landslides* **2023**, *20*, 1679–1689. [CrossRef]
39. Wang, D.; Gong, B.; Wang, L. On calibrating semantic segmentation models: Analyses and an algorithm. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 23652–23662.
40. Fernández, T.; Irigaray, C.; El Hamdouni, R.; Chacón, J. Methodology for landslide susceptibility mapping by means of a GIS. Application to the Contraviesa area (Granada, Spain). *Nat. Hazards* **2003**, *30*, 297–308. [CrossRef]
41. Wu, C.; Hu, K.; Liu, W.; Wang, H.; Hu, X.; Zhang, X. Morpho-sedimentary and stratigraphic characteristics of the 2000 Yigong River landslide dam outburst flood deposits, eastern Tibetan Plateau. *Geomorphology* **2020**, *367*, 107293. [CrossRef]

42. Chen, X.; Liu, C.; Chang, Z.; Zhou, Q. The relationship between the slope angle and the landslide size derived from limit equilibrium simulations. *Geomorphology* **2016**, *253*, 547–550. [CrossRef]
43. Çellek, S. Effect of the slope angle and its classification on landslide. *Nat. Hazards Earth Syst. Sci. Discuss.* **2020**, *2020*, 1–23.
44. Xiao, H.X.; Jiang, N.; Chen, X.Z.; Hao, M.H.; Zhou, J.W. Slope deformation detection using subpixel offset tracking and an unsupervised learning technique based on unmanned aerial vehicle photogrammetry data. *Geol. J.* **2023**, *58*, 2342–2352. [CrossRef]
45. Taylor, L.; Nitschke, G. Improving deep learning with generic data augmentation. In Proceedings of the 2018 IEEE Symposium Series on Computational Intelligence (SSCI), Bangalore, India, 18–21 November 2018; pp. 1542–1547.
46. Chlap, P.; Min, H.; Vandenberg, N.; Dowling, J.; Holloway, L.; Haworth, A. A review of medical image data augmentation techniques for deep learning applications. *J. Med. Imag. Radiat. Oncol.* **2021**, *65*, 545–563. [CrossRef] [PubMed]
47. Yang, J.; Tu, J.; Zhang, X.; Yu, S.; Zheng, X. TSE DeepLab: An efficient visual transformer for medical image segmentation. *Biomed. Signal Process. Control* **2023**, *80*, 104376. [CrossRef]
48. Qian, J.; Ci, J.; Tan, H.; Xu, W.; Jiao, Y.; Chen, P. Cloud detection method based on improved deeplabV3+ remote sensing image. *IEEE Access* **2024**, *12*, 9229–9242. [CrossRef]
49. Baheti, B.; Innani, S.; Gajre, S.; Talbar, S. Semantic scene segmentation in unstructured environment with modified DeepLabV3+. *Pattern Recognit. Lett.* **2020**, *138*, 223–229. [CrossRef]
50. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef]
51. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
52. Shi, W.; Meng, F.; Wu, Q. Segmentation quality evaluation based on multi-scale convolutional neural networks. In Proceedings of the 2017 IEEE Visual Communications and Image Processing (VCIP), St. Petersburg, FL, USA, 10–13 December 2017; pp. 1–4.
53. Yan, Z.; Yan, M.; Sun, H.; Fu, K.; Hong, J.; Sun, J.; Zhang, Y.; Sun, X. Cloud and cloud shadow detection using multilevel feature fused segmentation network. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 1600–1604. [CrossRef]
54. Al-Najjar, H.A.; Kalantar, B.; Pradhan, B.; Saeidi, V.; Halin, A.A.; Ueda, N.; Mansor, S. Land cover classification from fused DSM and UAV images using convolutional neural networks. *Remote Sens.* **2019**, *11*, 1461. [CrossRef]
55. Couprie, C.; Farabet, C.; Najman, L.; LeCun, Y. Indoor semantic segmentation using depth information. *arXiv* **2013**, arXiv:1301.3572.
56. Pena, J.; Tan, Y.; Boonpook, W. Semantic segmentation based remote sensing data fusion on crops detection. *J. Comput. Commun.* **2019**, *7*, 53–64. [CrossRef]
57. Sun, Y.; Zuo, W.; Yun, P.; Wang, H.; Liu, M. FuseSeg: Semantic segmentation of urban scenes based on RGB and thermal data fusion. *IEEE Trans. Autom. Sci. Eng.* **2020**, *18*, 1000–1011. [CrossRef]
58. Sun, Y.; Fu, Z.; Sun, C.; Hu, Y.; Zhang, S. Deep multimodal fusion network for semantic segmentation using remote sensing image and LiDAR data. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5404418. [CrossRef]
59. Wang, K.; He, D.; Sun, Q.; Yi, L.; Yuan, X.; Wang, Y. A novel network for semantic segmentation of landslide areas in remote sensing images with multi-branch and multi-scale fusion. *Appl. Soft Comput.* **2024**, *158*, 111542. [CrossRef]
60. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520.
61. Aufar, Y.; Kaloka, T.P. Robusta coffee leaf diseases detection based on MobileNetV2 model. *Int. J. Electr. Comput. Eng.* **2022**, *12*, 6675. [CrossRef]
62. Indraswari, R.; Rokhana, R.; Herulambang, W. Melanoma image classification based on MobileNetV2 network. *Procedia Comput. Sci.* **2022**, *197*, 198–207. [CrossRef]
63. Ahsan, M.M.; Nazim, R.; Siddique, Z.; Huebner, P. Detection of COVID-19 patients from CT scan and chest X-ray data using modified MobileNetV2 and LIME. *Healthcare* **2021**, *9*, 1099. [CrossRef]
64. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
65. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
66. Ma, N.; Zhang, X.; Zheng, H.; Sun, J. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 116–131.
67. Yu, H.; Chen, C.; Du, X.; Li, Y.; Rashwan, A.; Hou, L.; Jin, P.; Liu, F.F.; Kim, J.; Li, J. TensorFlow Model Garden. 2020. Available online: https://github.com/tensorflow/models (accessed on 21 December 2020).
68. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
69. 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.
70. Zhao, W.; Wang, R.; Liu, X.; Ju, N.; Xie, M. Field survey of a catastrophic high-speed long-runout landslide in Jichang Town, Shuicheng County, Guizhou, China, on July 23, 2019. *Landslides* **2020**, *17*, 1415–1427. [CrossRef]
71. Jongmans, D.; Garambois, S. Geophysical investigation of landslides: A review. *Bull. Société Géologique Fr.* **2007**, *178*, 101–112. [CrossRef]

72. Máttyus, G.; Luo, W.; Urtasun, R. Deeproadmapper: Extracting road topology from aerial images. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3438–3446.
73. Milletari, F.; Navab, N.; Ahmadi, S. 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; pp. 565–571.

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

Landslide Susceptibility Assessment in Active Tectonic Areas Using Machine Learning Algorithms

Tianjun Qi *, Xingmin Meng and Yan Zhao

School of Earth Sciences, Lanzhou University, Lanzhou 730000, China; xmmeng@lzu.edu.cn (X.M.); zhaoyandz@lzu.edu.cn (Y.Z.)
* Correspondence: qitj16@lzu.edu.cn

Abstract: The eastern margin of the Tibetan Plateau is one of the regions with the most severe landslide disasters on a global scale. With the intensification of seismic activity around the Tibetan Plateau and the increase in extreme rainfall events, the prevention of landslide disasters in the region is facing serious challenges. This article selects the Bailong River Basin located in this region as the research area, and the historical landslide data obtained from high-precision remote sensing image interpretation combined with field validation are used as the sample library. Using machine learning algorithms and data-driven landslide susceptibility assessment as the methods, 17 commonly used models and 17 important factors affecting the development of landslides are selected to carry out the susceptibility assessment. The results show that the BaggingClassifier model shows advantageous applicability in the region, and the landslide susceptibility distribution map of the Bailong River Basin was generated using this model. The results show that the road and population density are both high in very high and high susceptible areas, indicating that there is still a significant potential landslide risk in the basin. The quantitative evaluation of the main influencing factors emphasizes that distance to a road is the most important factor. However, due to the widespread utilization of ancient landslides by local residents for settlement and agricultural cultivation over hundreds of years, the vast majority of landslides are likely to have occurred prior to human settlement. Therefore, the importance of this factor may be overestimated, and the evaluation of the factors still needs to be dynamically examined in conjunction with the development history of the region. The five factors of NDVI, altitude, faults, average annual rainfall, and rivers have a secondary impact on landslide susceptibility. The research results have important significance for the susceptibility assessment of landslides in the complex environment of human–land interaction and for the construction of landslide disaster monitoring and early warning systems in the Bailong River Basin.

Keywords: landslides; susceptibility; machine learning; active tectonic areas; Bailong River Basin

1. Introduction

Globally, landslides remain the most dominant type of geological hazard in mountainous areas, causing a significant loss of life and damage to infrastructure every year [1,2], especially with intense seismic activity, extreme rainfall, and snowmelt events leading to an increasing trend of group-occurring landslides [3–5]. How to prevent landslide disasters with high efficiency and precision and low cost in vast mountainous areas has become an important challenge for the construction of geological disaster prevention systems in many countries. In the research on preventing regional landslide disasters, susceptibility assessment has become an important element that cannot be ignored and replaced [6]. Landslide susceptibility assessment is an important method for evaluating the possibility of landslide occurrence. By analogy with the environment in which landslides have occurred in the past, the likelihood of future landslides in each surveying unit (grid or slope unit) is evaluated [7]. The results have important practical value for disaster prevention planning, decision-making, and deployment of monitoring equipment within the

region [8]. The traditional assessment of landslide susceptibility mainly relies on on-site investigation, recording by researchers, and feedback from ground contact monitoring equipment [9]. Although the evaluation results of these methods are accurate and reliable, they are inefficient and require a great deal of manpower and financial resources, especially in high-altitude mountainous areas with complex terrain conditions, which heavily rely on geological surveys to obtain landslide inventories. It is unrealistic to conduct a landslide susceptibility assessment based on investigation and monitoring in mountainous areas ranging from hundreds to tens of thousands of square kilometers. This phenomenon has been completely improved with the popularization of high-precision remote sensing image, DEM (Digital Elevation Model), geological map, rainfall, and other refined data, as well as the significant improvements in software and technical performance, such as Geographical Information System platforms [10–12].

In recent years, methods of landslide susceptibility assessment driven by big data have significantly taken a dominant position and begun to play an increasingly important role. In addition to the development of air–space–ground remote sensing technology that provides a large amount of high-quality geographic data for landslide disaster analysis, end-to-end machine/deep learning technology can automatically fit labeled datasets without the need for in-depth research on complex landslide mechanisms to achieve prediction of potential locations of landslide occurrence [13]. Models for evaluating landslide susceptibility have emerged in large numbers with the deepening of research [14,15]. However, it is widely believed that different models have adaptability issues for the environment and scale [16,17]. Therefore, selecting advantageous models suitable for regional features from a large number of models has become a key step in conducting susceptibility assessments in complex regions. The rapid development of artificial intelligence (AI) technology in recent years has provided broad prospects for solving this issue and has achieved success in various geological hazard assessments [14,18,19]. In addition, the advantage of using machine learning for landslide susceptibility assessment is that it can quantitatively reveal the relative contribution of different factors to landslide development among numerous influencing factors [12,13]. The significance of regional disaster prevention lies precisely in the targeted prevention and control of landslide development based on specific factors, which is also the most important value of landslide susceptibility assessment.

The eastern margin of the Tibetan Plateau is a globally renowned zone of rapid terrain change, with active tectonic movements, strong river incision and downcutting, and high population density [1,2], which makes it one of the most severe areas for landslide disasters in the world. The complex geological environment, extremely variable climate, and excessive human activities have caused landslides in the region to be characterized by suddenness, group occurrence, and multi-type concurrency, making prediction and management extremely difficult. Due to various factors such as terrain conditions, vegetation coverage, and disturbance by human activities, it is difficult to systematically catalog and evaluate the susceptibility of landslides in this area. In order to effectively mitigate the risk of landslide disasters and carry out systematic monitoring in the key areas, this article selects the Bailong River Basin on the eastern margin of the Tibetan Plateau as the research area. Using machine learning methods, the optimal model for evaluating susceptibility in alpine-gorge areas is selected from various algorithms, and quantitative evaluation of the main factors that affect landslide development is attempted. The final results can provide scientific reference for the monitoring and early warning of landslides in the complex human–land interaction environment on the eastern margin of the Tibetan Plateau.

2. Study Area

The Bailong River is located at the intersection of the eastern margin of the Tibetan Plateau and the West Qinling Mountains (Figure 1) and is a secondary tributary of the Yangtze River and a primary tributary of the Jialing River. With the development of large-scale fault zones and intense topographic relief, the Bailong River Basin has become one of the most serious regions for landslide disasters in China under the combined effects of

earthquakes, heavy rainfall, and human engineering activities [20,21]. The total population within the watershed exceeds one million, mainly distributed in the valley and hillslope zones on both sides of the Bailong River. Long-term tracking investigations have found that due to high population density and a scarcity of land resources, the widespread utilization of old landslides as buildings, roads, and agricultural land is one of the main reasons for serious landslide disasters. The risk of landslide reactivation forming is a serious threat to the lives and properties of local residents. For example, on 12 July 2018, a large-scale landslide in Nanyu Township, Zhouqu County was reactivated, and the landslide deposit blocked most of the Bailong River channel to form a dam and backwater, causing the water level of the Bailong River to rise by 8 m in a short period of time, flooding the bridges, roads, hydropower stations, and most residential buildings in Nanyu Township [22]; and on 19 July 2019, the Yahuokou landslide, which was a long strip-shaped landslide formed along a groove of the fault zone, reactivated with a length of up to 2.0 km and an average width of less than 100 m. After the landslide occurred, it destroyed the village roads and the factory buildings at the lower edge of the slope that were perched on the old landslide body [23]; on 16 August 2020, a continuous heavy rainfall event in the Bailong River Basin triggered a large number of shallow landslides, causing serious economic losses and ecological damage; and on 18 January 2021, the Lijie landslide in Beishan Village, Zhouqu County, was reactivated, which led to the emergency evacuation of thousands of people in the township [24]. The above cases are only the tip of the iceberg of landslide disasters in the Bailong River Basin.

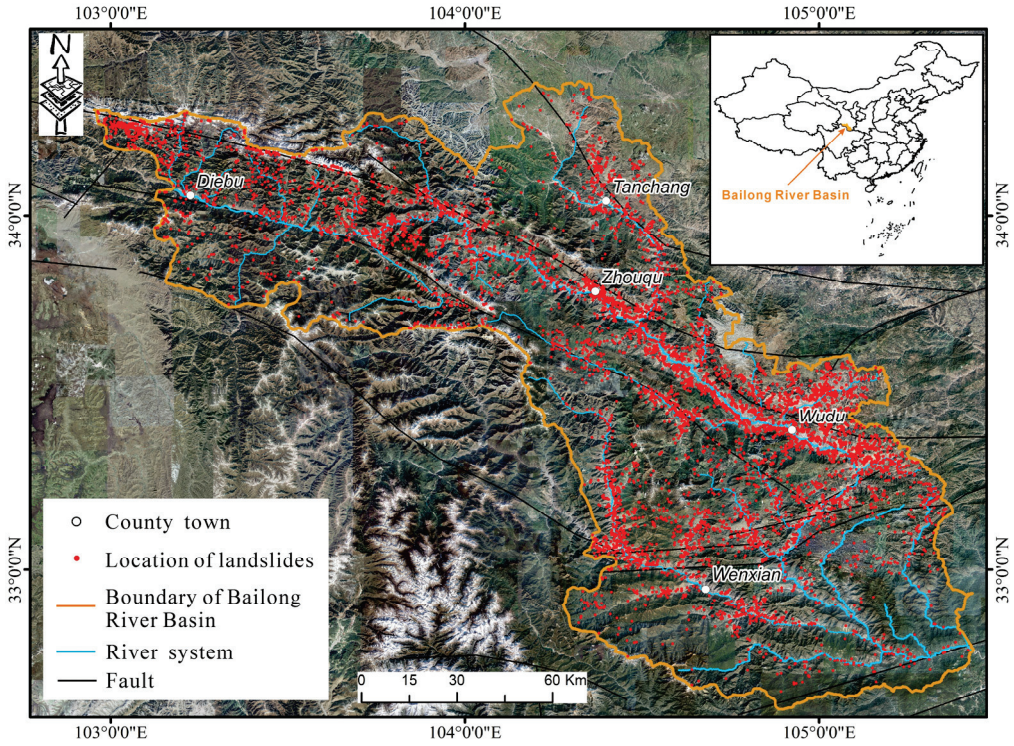


Figure 1. The distribution of historical landslides in the Bailong River Basin.

The serious geological disasters that occur in the Bailong River Basin arise first from the geological structural conditions. Driven by the far-field stress caused by the expansion of the Tibetan Plateau, the collision between the Bailongjiang Block and the Bikou Block formed structural characteristics of the basin [25], with a stratigraphic structure mainly composed of Carboniferous to Triassic limestone nappes and lower Silurian weak phyllite. Since 1.7 Ma, the rapid downcutting of rivers has led to the abundant exposure of weak phyllite and carbonaceous slate. As a result, earthquakes and rainfall have become the main triggering factors for landslides in the limestone nappe and phyllite areas, respectively.

The dry–hot valley climate within a local area is another important reason for the frequent occurrence of landslides. Since the 1950s, extensive deforestation has directly contributed to the exacerbation of the dry–hot valley climate in the basin, with frequent occurrences of extreme rainfall events. Therefore, under the coupling and synergistic effect of internal and external factors, this area has become an ideal site for landslide development.

Despite the extreme development of historical landslides in the Bailong River Basin, a systematic landslide inventory is still lacking. The previous records of landslides in this area mainly include sites of landslides on both sides of highways and river channels that pose a threat; however, there is a lack of specific parameter indicators such as range and area, making it difficult to use these records for susceptibility assessment. Historical landslide inventories are an important foundation for studying spatial distribution patterns, susceptibility assessment, risk evaluation, and formation mechanisms [26]. Qi utilized the popularization of high-resolution remote sensing images and compiled a historical landslide database of the entire watershed for the first time through remote sensing interpretation combined with on-site verification, as shown in Figure 1 [27]. This inventory covers 6609 landslides in the Bailong River Basin, including 140 giant landslides, 2619 large landslides, 3176 medium-sized landslides, and 674 small landslides. Visual inspection reveals that landslides in this area are mainly distributed along fault zones and on both sides of river valleys, closely related to the scope of human distribution and activity. Therefore, it is urgent to conduct a landslide susceptibility assessment in this area to provide a foundation for systematic disaster prevention and reduction planning.

3. Methods and Materials

3.1. Selection of Evaluation Factors

The reason why the complex geological environment and the scope of human activities easily become the ideal development areas for landslides is that these areas contain almost all the necessary factors that induce and affect the occurrence of landslides. The Bailong River Basin is located at the junction of the eastern margin of the Tibetan Plateau and the West Qinling Mountains. It is the transition zone of China's first- and second-order terrain steps. During the formation of the Qinling Mountains in the Indosinian and the uplift of the Tibetan Plateau in the Cenozoic, it experienced strong crustal uplift and thrust compression, resulting in the formation of a wide distribution of fault fracture zones, accompanied by the erosion of rivers, resulting in the formation of a landscape with ridges and valleys. The region has a typical mountain climate, and extreme rainfall events are the main external factors inducing landslides. In recent years, intensive human activities have had a profound impact on the formation and development of landslides in the study area. Based on this background, the landslide control and inducing factors considered in this paper involve a wide range of fields, including terrain (slope, aspect, altitude, local relief, plane curvature, profile curvature, topographic wetness index, surface roughness, water system), geological structure (lithology, fault layer, topographic/bedding-plane intersection angle (TOBIA) [28]), material conditions (land use, soil type, Normalized Difference Vegetation Index (NDVI)), and human activities and rainfall (road, population density, annual rainfall). Each factor may become a direct or indirect inducing condition of landslides. Table 1 briefly describes the types and sources of these factors. The extraction of geomorphic factors is based on DEM data with a resolution of 12.5 m (from Japan Aerospace Exploration Agency's ALOS satellite); the geological factors are derived from the 1:50,000 and 1:100,000 geological maps of China; the soil data and land use data are from the global land-cover product with fine classification system at 30 m using time-series Landsat imagery [29]; NDVI data is from GF-1 8 m resolution satellite data; road data is obtained by interpreting Google Earth image with 1 m resolution; and rainfall data are from the monitoring data of rainfall stations in Bailong River Basin.

Table 1. Fields and characteristics of the spatial database.

Type	Data	Abbr.	Source	Spatial Resolution
Topography	Slope gradient	AS	Derived from DEM	12.5 m
	Slope aspect	SA	Derived from DEM	12.5 m
	Elevation	EL	Derived from DEM	12.5 m
	Local relief	LR	Derived from DEM	12.5 m
	Planar curvature	PLC	Derived from DEM	12.5 m
	Profile curvature	PRC	Derived from DEM	12.5 m
	Topographic wetness index	TWI	Derived from DEM	12.5 m
	Surface roughness index	SR	Derived from DEM	12.5 m
	Distance to river	DR	Derived from DEM	12.5 m
Geology	Formation lithological index	FLI	Geo-map	1:100,000
	Distance to fault	DF	Geo-map	1:50,000
	Topographic/bedding-plane intersection angle	TOBIA	Geo-map	1:100,000
Material	Land use	LU	GLC_FCS30-2020	30 m
	Soil type	ST	HWSD	1 km
	Normalized Difference Vegetation Index	NDVI	GF-1 satellite	8 m
Human activity and inducing factors	Distance to road	DTR	Google Earth image	1 m
	population density	PD	91 Weitu	1 km^2
	Annual precipitation index	API	2000–2020	year

3.2. Data Preprocessing

Among the selected factors, some of them may have strong correlation to a certain extent, showing an approximate linear relationship, which will disrupt the regression and fitting process of the model and affect the stability of the model operation [30]. The strong correlation of factors has an impact on the regression and fitting process: the regression and fitting of the model aims to find the relationship between each factor and the target variable. Assuming that the correlation between the two factors is too close, the change of one factor may lead to the corresponding change of the other factor when dealing with the conditional factor and the target variable, so it is impossible to fix other conditions to analyze the impact of a single factor on the output result, because one factor in the fitting will be mixed with the impact of another factor. Therefore, collinearity analysis is the first step in regression and fitting.

3.3. Selection of Machine Learning Algorithms

With the rapid development of artificial intelligence technologies, machine learning has flourished in various industries since the 1990s with significant achievements, including the implementation of neural networks, the development of boosting algorithms, and increased accessibility to internet-derived and digital data. Machine learning was first used in the field of landslides in the early to mid-2000s, where machine learning modeling demonstrated strong superiority in terms of arithmetic power and predictive accuracy compared to traditional physical-based models, heuristic models, and statistical models, with many machine learning models having predictive accuracies of 90% or more. A lot of popular machine learning algorithmic architectures appeared in the following years, including Logistic Regression, Support Vector Machine, Decision Tree, Random Forest, Bernoulli-NB and Gaussian-NB, Bagging Classifier, K-Nearest Neighbors, and Gradient Boosting algorithms, etc. [19]. However, to date, there is no consensus on which machine learning algorithms are "best suited" to predict landslide susceptibility areas. Many studies have shown that the prediction accuracy of data-driven landslide susceptibility modeling is affected not only by the quality of the data of landslide inventory and landslide condition factors but also by the underlying quality of the machine learning algorithms used [31].

Therefore, in order to compare the adaptability of different models in complex geological environments, we selected 17 models that are most widely used in landslide susceptibility assessment to screen the most suitable models for this study area. The characteristics of these algorithms are shown in Table 2.

Table 2. Selected algorithms and characteristics.

Models	Characteristic
Random Forest Classifier	Uses many classification trees to stabilize model predictions. Each decision of a tree is further based on a randomly selected predictor, and the predictions of category assignments are determined by a majority vote of all trees. The proportion of trees predicting the existence of landslides in the set can be used as an indicator of landslide susceptibility.
Bagging Classifier	An ensemble algorithm that establishes multiple instances estimated by black boxes on a random subset of the original training set and then, aggregates these predictions to form the final prediction.
K-Neighbors Classifier	For the training set, the categories of the individual instances have been determined. During the classification process, for new instances, predictions are made through majority voting based on the categories of their K nearest neighbor training instances.
Decision Tree	An instance-based inductive learning method that can refine a tree-like classification model from a given unordered training sample.
Extra Tree	A variant of Random Forest.
Gradient Boosting	This model applies the gradient descent technique to the regression tree. The principle is to treat the value of the basic learner (regression tree) in each iteration on x as the negative gradient of the loss function space on x, and the coefficient before the basic learner is treated as the step size to approximate the minimum value of the error function space.
XGBoost	Uses the boosting technique to randomly divide the initial sample set into k parts, and then divides each subset into a training set and a validation set by a 2:1 ratio to generate a decision tree.
AdaBoost	An integrated learning technology that can turn a weak learner into a strong learner with higher prediction accuracy.
Logistic Regression	A fitting method for classifying records based on the values of conditional variables to estimate the probability of an event occurring.
Linear Discriminant Analysis	Involves the projection of high-dimensional pattern samples into the space of the best discriminating vectors to extract categorical information and compress the dimensionality of the feature space.
SGDClassifier	Achieved in a “one-vs-all (OVA)” manner by combining multiple binary classifications.
Bernoulli-NB and Gaussian-NB	Based on the concept of Bayesian probability, assuming that each attribute is independent of all other attributes to obtain the probability of each feature, and using a higher probability as the prediction result.
Quadratic Discriminant Analysis	Here, the assumptions made are more stringent than those of Logistic Regression, but when these assumptions are met, discriminant analysis can be used as a useful alternative or supplement to Logistic Regression.
Passive Aggressive Classifier	An online learning algorithm used for regression and classification. Compared to Support Vector Machine, it is easy to use and works faster, but cannot provide high accuracy like Support Vector Machine.
Perceptron	A supervised learning algorithm based on binary classification that can predict whether the input represented by a digital vector belongs to a specific class.

3.4. Parameter Preprocessing

The assessment of landslide susceptibility prediction results is mainly used to test the performance and operation of landslide susceptibility prediction models subjected to various uncertainty issues and is a very important step in the modeling of landslide susceptibility prediction [3]. The results are evaluated in order to prove the validity of the study; therefore, almost every paper includes a performance evaluation of landslide susceptibility prediction. Currently, there are numerous evaluation parameters in susceptibility modeling, and the widely used evaluation parameters include the following: i. ACC, which indicates the rate of correctness of all samples involved in the modeling; ii. PPV, which is the proportion of correctly predicted positive instances to the proportion of all predictions that are positive; iii. TPR, which indicates the rate of recall or the rate of checking for completeness; iv. TNR, indicating the ability to predict negative instances; v. AUC, indicating the area under the receiver operating characteristic curve (ROC); the value indicates the degree of fit and reflects the balance between sensitivity and specificity; and vi. TSA, which indicates the correctness rate of the samples that do not participate in model training, reflecting the generalization of the model. Of the above, ACC is the most widely cited. This paper also uses this method to evaluate the model.

$$\mathrm{ACC} = (\mathrm{TP} + \mathrm{TN})/(\mathrm{TP} + \mathrm{FN} + \mathrm{FP} + \mathrm{TN})$$

These terms are listed in Table 3 and are defined as follows:

True positive (TP): the predicted class is positive, and the prediction is consistent with the actual situation;

False positive (FP): the predicted class is positive, and the prediction is opposite to the actual situation;

True negative (TN): the predicted class is negative, and the prediction is consistent with the actual situation;

False negative (FN): the predicted class is negative, and the prediction is opposite to the actual situation.

Table 3. Confusion matrix results.

		Predicted Label	
		Positive	**Negative**
True label	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

4. Results

4.1. Collinearity Analysis of Factors

In this paper, the collinearity analysis of machine learning is used to calculate the heat map of the correlation matrix of the selected parameters (Figure 2), which is based on Seaborn Python 2.7, https://seaborn.pydata.org/generated/seaborn.heatmap.html#seaborn.heatmap (accessed on 15 October 2021). The main purpose of visual calculation is to eliminate the collinearity problem of the selected parameters. In the matrix heat map, it can be found that there is a high correlation between SR and LR, and we removed the parameters with a correlation coefficient > 0.7, as proposed by Dormann et al. [30]. Therefore, in this study, we exclude the factor of surface roughness index (SR) and select the other 17 factors for model screening.

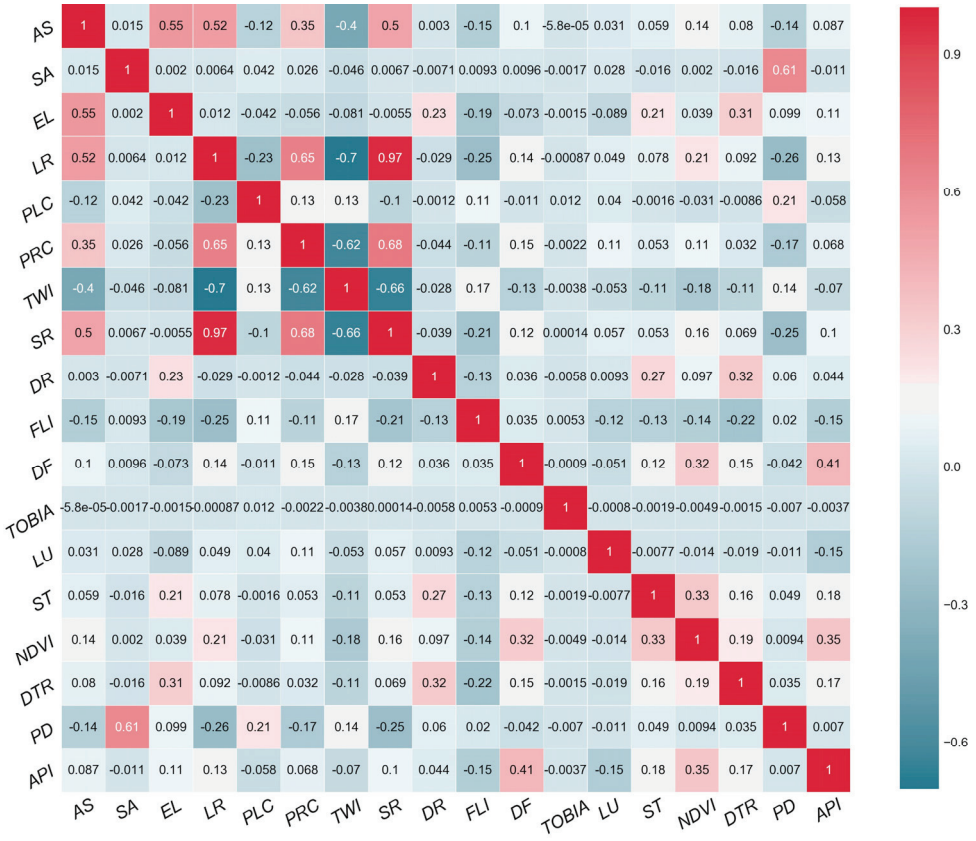


Figure 2. Correlation heatmap of influencing factors.

4.2. Evaluation and Optimization of Models

The initial model is trained using the data in the cross-validation training data set, and then, the results of the models are ranked according to the average accuracy score (ACC) of the test data in the cross-validation dataset. Finally, the model evaluation score is generated, as shown in Figure 3. Among the evaluation results of all models, the results of the ensemble models are better than those of other models, and the calculation results of all models are greater than 0.6. Among these, RandomForestClassifier has the highest score of 0.9, followed by BaggingClassifer at 0.87.

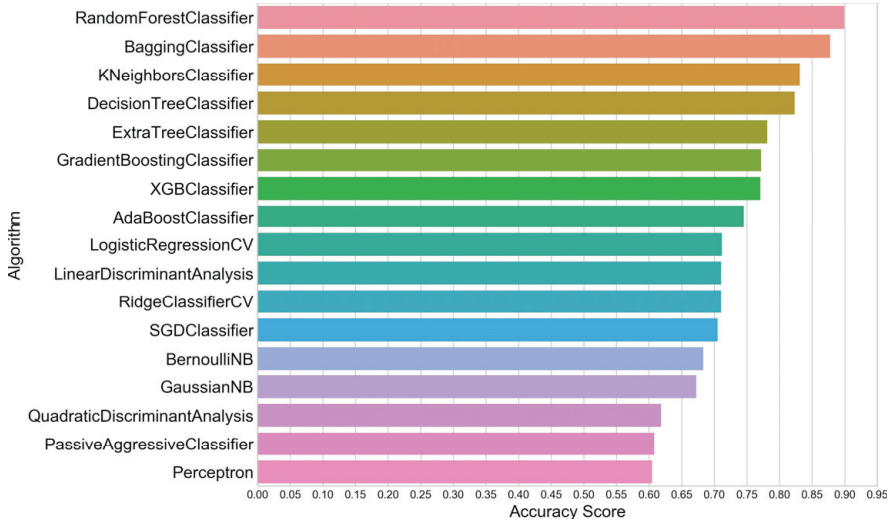


Figure 3. Ranking of model accuracy scores.

In order to improve the accuracy of these models, this paper chooses to optimize the model with the top four scores in the initial prediction. The optimization method uses the parameter grid method and grid search cross-validation method to fit the model, find the best super parameter through the ACC score, cross-validate the training set of the model 10 times according to the optimal super parameter of each model (Figure 4), and reorder the model according to the average accuracy score of the test data. After optimization, the performance of all four models has been improved (Figure 5). BaggingClassifier has the largest improvement, surpassing RandomforestClassifier as the best model. For BaggingClassifier, the ACC test data of 10 cross-verifications is 0.91, and the average AUC is 0.96. The performance of the other three models has also improved after optimization, but the extent of improvement is relatively small. After evaluation, the predicted probability of the BaggingClassifier model was finally chosen to map the landslide susceptibility of the Bailong River Basin.

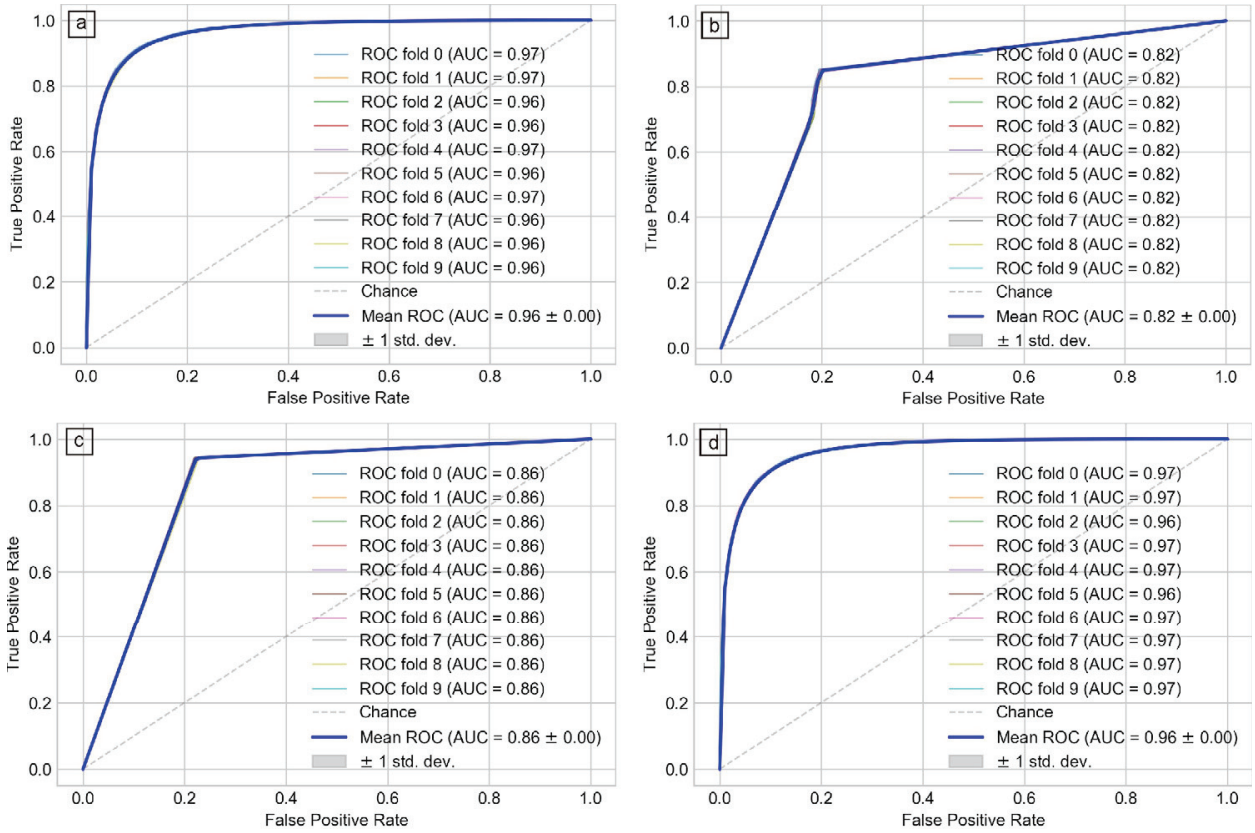


Figure 4. ROC curve and AUC value of four models with 10 cross-validations, (**a**) BaggingClassifier; (**b**) DecisionTreeClassifer; (**c**) KNeighborsClassifer; (**d**) RandomForestClassifier.

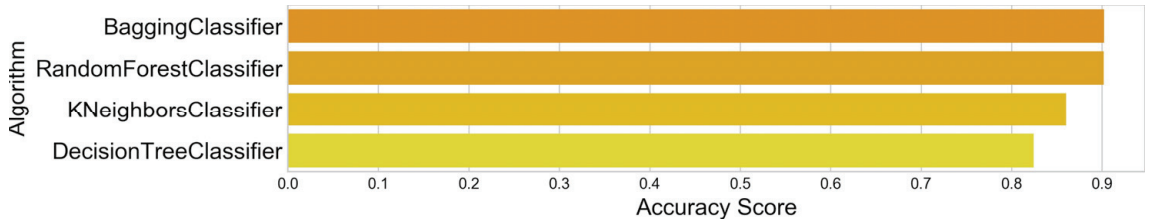


Figure 5. Ranking of ACC scores after model optimization.

4.3. Landslide Susceptibility Mapping

The results of landslide susceptibility assessment in the Bailong River Basin predicted by the BaggingClassifier model are shown in Figure 6. The prediction result of susceptibility is a probability value between 0 and 1. According to the classification principle of the natural discontinuity method, the susceptibility result is divided into five intervals: very low (0–0.08), low [0.08–0.19), moderate [0.19–0.39), high [0.39–0.88), and very high [0.88–1). The proportion of area occupied by each susceptibility zone within the watershed varies. According to the statistics of the area and proportion of susceptibility areas, the area ratio of the first four types of susceptibility gradually decreases in proportion to the area as the level increases. The very low susceptibility areas accounted for more than half of the total area of the region, accounting for 52.86%, and the low susceptibility areas accounted for 21.94%. The moderate and high susceptibility areas accounted for the lowest proportion in the whole basin, accounting for 8.32% and 5.88%, respectively. The area of very high susceptibility areas reached 2009.6 km^2, accounting for 11% of the total basin area. The distribution of susceptibility in the basin shows the characteristics of a large proportion of low and moderate susceptibility areas and a concentrated distribution of high susceptibility areas.

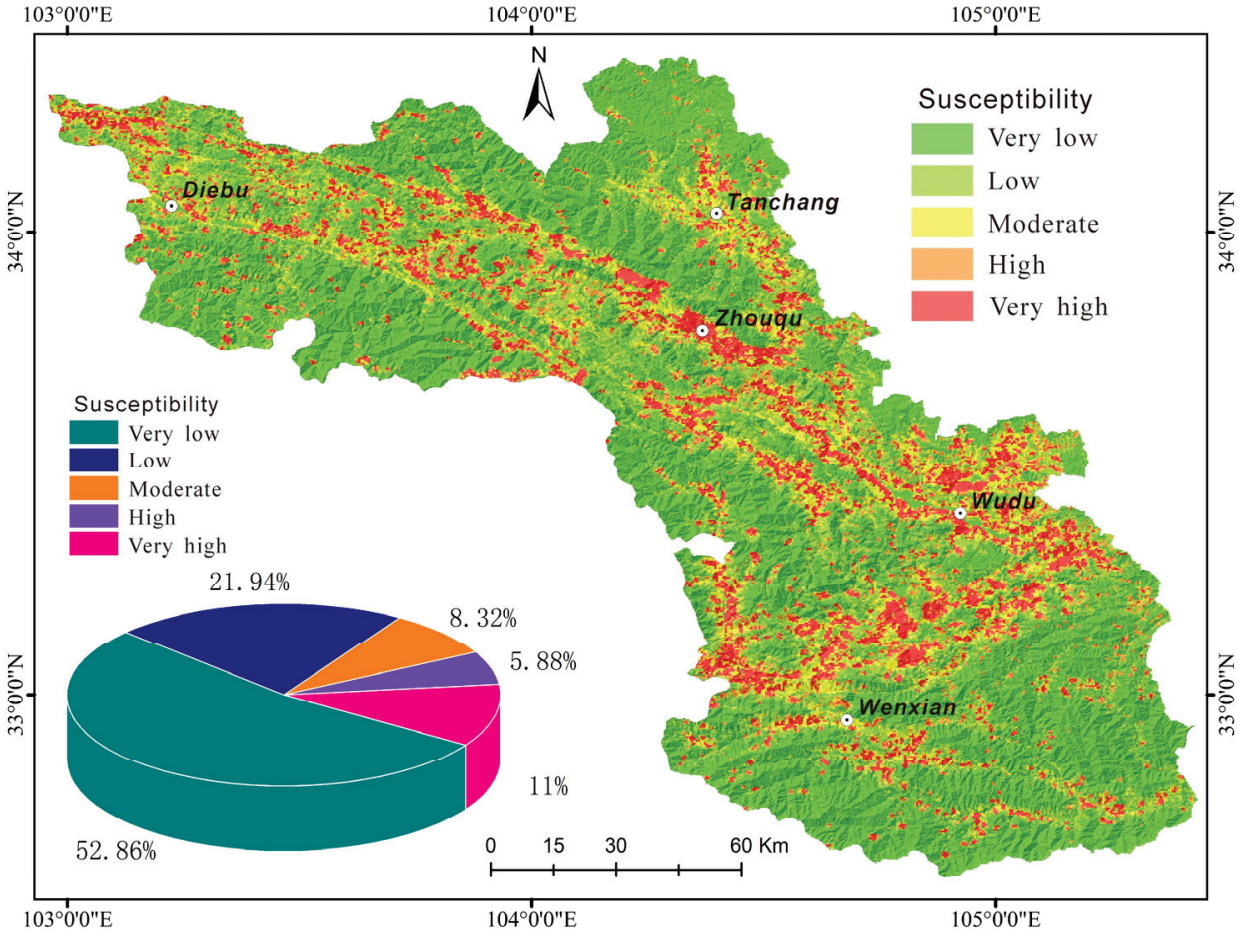


Figure 6. Distribution map of landslide susceptibility assessment in Bailong River Basin.

The prediction results of landslide susceptibility in the Bailong River Basin show that the low susceptibility areas in the basin are mainly located in the north and south areas of Tanchang, the south of Diebu, and the south of Wenxian, which are mainly areas with high vegetation coverage and sparse distribution of faults. The vegetation has a certain protective effect on the surface. The distribution range of high and very high landslide susceptibility areas is closely related to the distribution of river systems and faults in the basin. In the upstream area of Wudu, the very high susceptibility areas are mainly

distributed along the main course and primary tributaries of the Bailong River. In the upstream of the Bailong River, very high susceptible areas are mainly distributed along the banks of the main course and primary tributaries of the Bailong River. In the downstream of Wudu, the high and very high susceptibility areas at the bend of the Bailong River are still distributed along both sides of the river. However, there are also large areas of very high susceptibility areas in the area north of Wenxian. Overall, the distribution of high and very high susceptibility areas in the Bailong River Basin are closely related to the fault zones and river system. In terms of spatial distribution, the high and very high susceptibility areas are closely related to the distribution of high population density and high road density areas in the Bailong River Basin (Figure 7), which poses a great challenge to disaster prevention and control in the basin. The susceptibility assessment results in this paper can be used as reference data for decision-makers and government managers in land management and allocation.

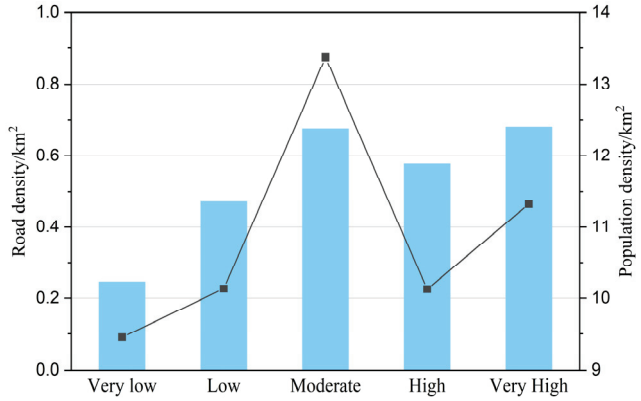


Figure 7. Statistical relationship between road density and population density in different susceptibility areas.

4.4. On-Site Verification of Susceptibility Assessment Results

Under the existing conditions, field survey is one of the auxiliary means to verify the landslide susceptibility assessment results, and the accuracy of the evaluation results can be judged intuitively through field survey. The high and very high susceptibility areas of the evaluation results are the focus of our attention and are also the key areas for future disaster prevention, land planning, and management. Therefore, we chose to investigate and verify the high and very high susceptibility zones of this susceptibility assessment result.

Considering that the study area is mainly controlled by fracture zones, whether the evaluation results can perform well along the fracture zones is an important test criterion. Therefore, in this paper, we choose the Zhouqu area located in the middle reaches of the Bailong River Basin as a susceptibility validation area (Figure 8a). This area is controlled by fracture zones with dramatic topographic relief, broken rock units, and extremely developed landslides, and has experienced serious landslide disasters in the historical period. The high and very high susceptibility areas in the susceptibility assessment results include landslides that have occurred and slopes with destabilization potential, and the validation of the two types of slopes shows that serious landslides have occurred in almost all of the high and very high susceptibility zones and that famous landslides in the study area, such as Suoertou landslide (Figure 8c), Xieliupo landslide (Figure 8d), Yahuokou landslide (Figure 8e), Lijie landslide (Figure 8i), etc., are in the very high susceptibility zones, and they are accompanied by rock fragmentation and local destabilization. This indicates that the susceptibility evaluation results obtained in this paper using machine learning are in line with the actual situation, have high credibility, and can be used as reference data for future disaster prevention and mitigation planning and land management.

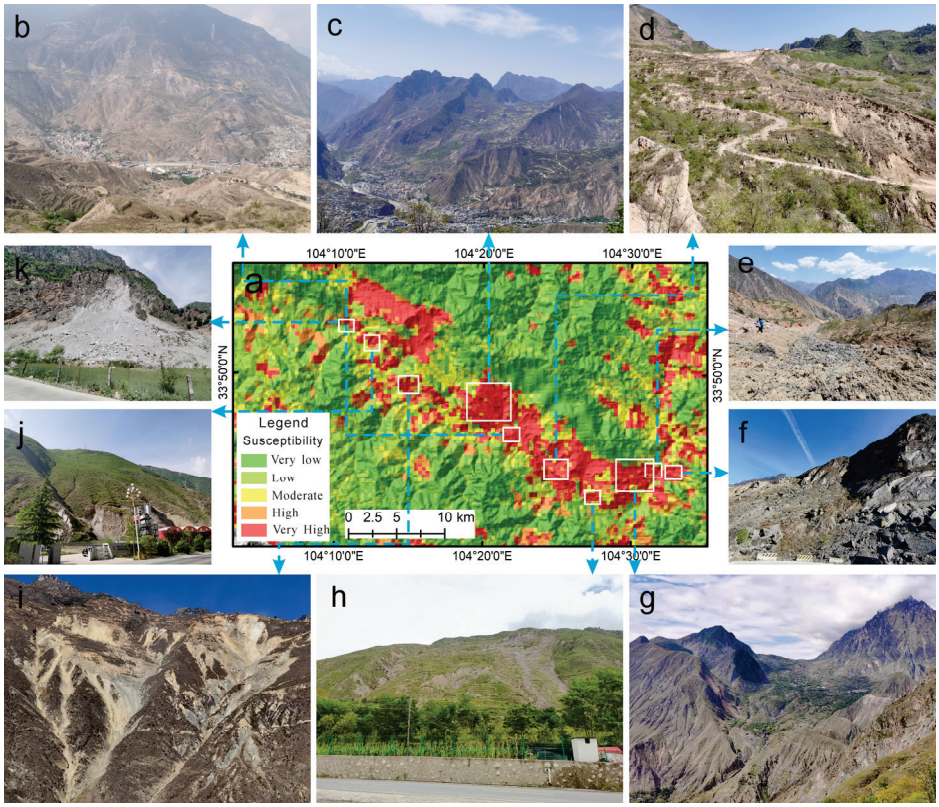


Figure 8. Field verification of landslide susceptibility assessment results. (**a**) Susceptibility assessment verification area, (**b**) Nanqiao lanslide, (**c**) Suoertou and Daxiaowan landslides, (**d**) Xieliupo landslide, (**e**) Yahuokou landslide, (**f**) A small scale rockfall, (**g**) Zhongpai landslide, (**h**) An old landslide, (**i**) Lijie landslide, (**j**) An old landslide, (**k**) A small scale rockfall.

5. Discussion

One of the advantages of machine learning modeling is that the interpretability of the model can be used to quantitatively evaluate the importance of the selected factors [13,32]. In this way, we can better understand the importance of various factors in landslide development and thus, take targeted prevention and control measures. The calculated importance weights of all factors are shown in Figure 9. There is no doubt that each factor plays a role in landslide development. Among all the factors, the distance to a road has the highest degree of impact, while the impact of land use type has the lowest effect. The impact of engineering construction on landslide development is unexpected. The impact of human activities on geological hazards in the Bailong River Basin has always been a concern for researchers. Engineering excavation will cause the readjustment of slope stress structures, which will seriously affect the stability of slopes [7,33]. In many studies of landslide susceptibility assessment, despite the repeated emphasis on the important influence of human activities, roads are not viewed as the most critical influence factor compared to the importance of condition factors such as slope, elevation, rainfall, and lithology [3,34], which is significantly different from the results of this paper. However, if the background of the Bailong River Basin is understood, roads may be acceptable as the factor most closely associated with landslide development. We regard that the importance of roads may be overestimated in the landslide susceptibility assessment of this area, and we did not consider that the dynamic process of the factors may be the reason for the overestimation or underestimation of the assessment results of some factors. Old and ancient landslides in the Bailong River Basin account for 73% of the total number of landslides, and the formation time of these landslides may be far older than the activity history in this region, which dates from the beginning of the human industrial age. Therefore, the formation of these old and ancient landslides is not the result of human

activities. As for the high value for the influence of DTR, we believe that the main reason for this is that the combination of complex topography, limited land resources, and rapid population expansion forced most residents to migrate to and settle on high hillslope areas, and these places of settlement and production activities are generally sites of large-scale ancient landslides and old landslide deposits in the historical period. In recent years, almost all villages have built hardened roads. Therefore, the importance of DTR in the factor evaluation is actually a feedback effect of landslide-based geomorphological processes on human activities, and landslide hazards bring risks and scarce land resources at the same time. Therefore, how to consider the dynamic connection between the historical landslide formation process and the condition factor is a key component to continuing to improve the reliability of the assessment results in the future. Nevertheless, the results of this paper are still of great revelation, because large landslides in the Bailong River Basin are predominantly ancient landslides and old landslides, and there is a potential risk of landslide reactivation under the influence of human activities, rainfall, and other factors. In fact, the reactivation of landslides has already caused serious consequences, and the likelihood of the occurrence of risk is even greater considering the close correlation that exists between roads and landslides. Therefore, there is a need for a systematic assessment and management of landslide and road relationships in order to avoid and mitigate the potential risks created by landslide utilization.

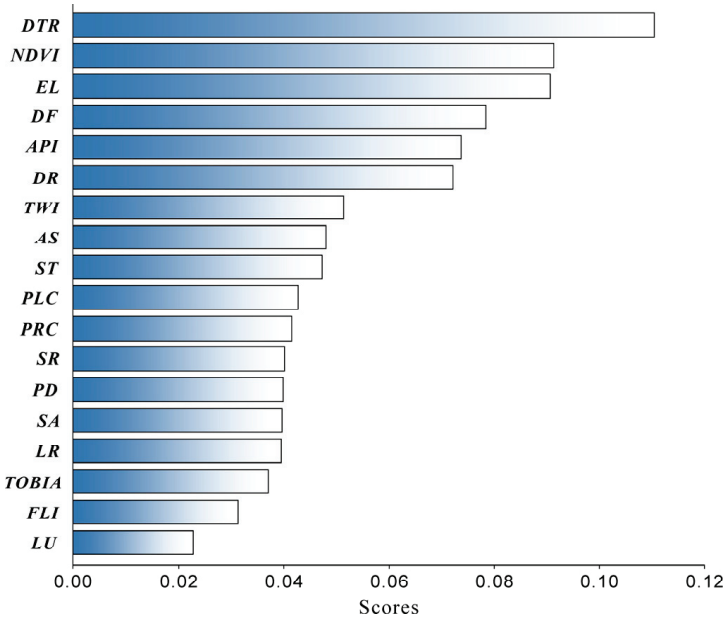


Figure 9. Quantitative evaluation results of importance of influencing factors.

NDVI is the second most important factor affecting landslide susceptibility with a 9% contribution. The statistical results show that in areas with vegetation coverage of less than 50% (Figure 10b), the proportion of landslide area is higher than that of non-landslide areas, while in areas with vegetation coverage of more than 50%, the proportion of non-landslide areas is significantly higher than that of landslide areas. This result clearly demonstrates the constraining effect of vegetation in the process of landslide development. The grid structure formed by vegetation roots is particularly effective in controlling shallow landslides [35]. Considering the explosive growth of geological disasters caused by human damage to forest resources and the ecological environment in the region in recent decades, ecological environment treatment and returning farmland to forest can be considered the priority choice for constraining geological disasters in the Bailong River Basin. The influence of faults on landslide development ranked fourth, accounting for 8% of the weight. It should be specified that the influence of fault zones on landslide development cannot be completely replaced by faults, including the influence of lithological fracture

and crustal uplift. From the analysis of the results, it can be seen that the proportion of landslide areas is significantly higher than that of non-landslide areas within 2000 m of a fault (Figure 10d). The influence of annual rainfall on the distribution of landslides is special. The distribution of landslides is concentrated in areas with annual rainfall in the range of 300 mm to 700 mm. Paradoxically, the proportion of landslides in low-rainfall areas is higher than that in high-rainfall areas (Figure 10e). There are three main reasons for this phenomenon: i. the rainfall in the Bailong River Basin is greatest in the Bikou area in the south and Diebu area in the north, however, as a stable craton block, the Bikou block experiences no violent tectonic activity, and very few landslides develop there, whereas the Diebu region has mostly tectonically formed ancient landslides with a lower density of landslides; ii. ecological damage reduces the stability threshold of the hillslope, which can induce a landslide under the condition of low rainfall, and ecological damage in the Bailong River Basin is concentrated in the area from Zhouqu to Wudu; thus the critical threshold of landslide occurrence in this area is low; iii. the section from Zhouqu to Wudu within the Bailong River Basin has a dry–hot valley climate, with concentrated rainfall, conditions that are more likely to induce landslides. However, it must also be acknowledged that the low resolution of rainfall data may also be the reason for this phenomenon. The influence of distances to a river is similar to that of distance to a fault in that the range of within 2000 m of a river is a high incidence area for landslides (Figure 10f), and in addition to erosion of the foot of the slope, the transport of material is necessary for the continuation of landslides.

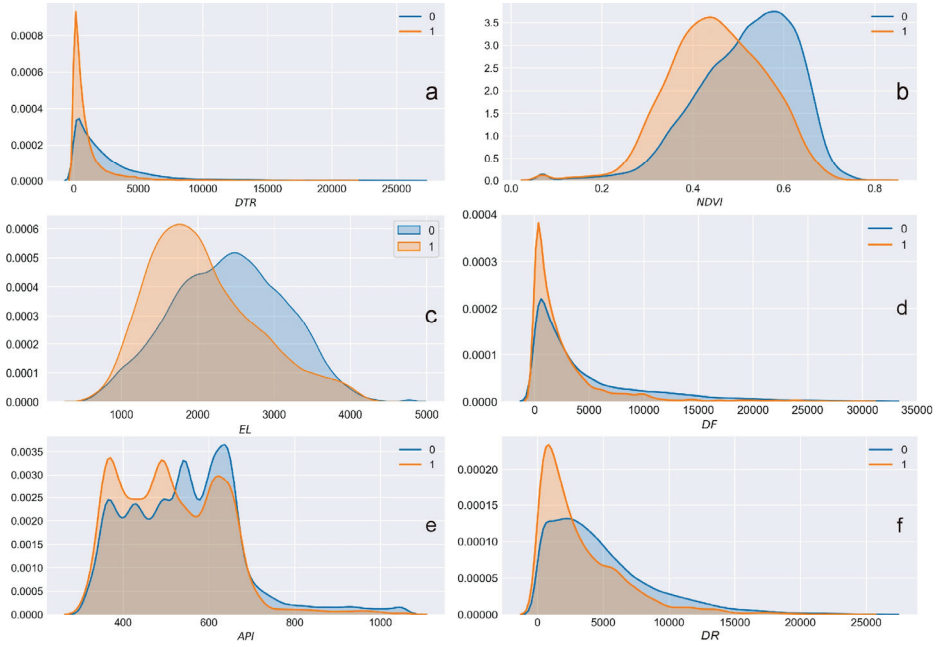


Figure 10. Landslide and non-landslide area distribution in some factors, (**a**) road, (**b**) NDVI, (**c**) altitude, (**d**) fault, (**e**) annual precipitation index, and (**f**) river. "1" represents landslide area and "0" represents non-landslide area.

6. Conclusions

The serious geological disasters that have occurred in the eastern margin of the Tibetan Plateau have made government managers and decision-makers face great challenges in urban planning, land management, ecological governance, and disaster prevention. The susceptibility assessment of landslide disasters as a quantitative and effective analysis result has a certain reference value. As the representative achievements of the 21st-century technological revolution, machine learning and artificial intelligence are becoming the dominant technologies leading the progress of human civilization. Artificial intelligence has shown great advantages in all professions. How to apply artificial intelligence in the field of geological disaster pre-

vention has always been a problem for practitioners. In this paper, the Bailong River Basin on the eastern margin of the Tibetan Plateau is taken as a typical study area, and 17 kinds of machine learning algorithms are selected for susceptibility assessment modeling, evaluation, and optimization. The results show that the BaggingClassifier and RandomforestClassifier models perform well, and the accuracy of validation is more than 90%. Using BaggingClassifier to generate predictions of Bailong River Basin landslide susceptibility, the findings are applied to divide the basin into very low, low, moderate, high, and very high susceptibility areas, accounting for 52.86%, 21.94%, 8.32%, 5.88%, and 11% of the basin, respectively. The high and very high landslide susceptibility areas are mainly distributed in the middle and lower reaches of the Bailong River from Zhouqu to Wenxian, close to the river banks, and are areas with high population density and road density in the basin, which indicates that the risk of landslide in the Bailong River Basin is high, and there is the possibility of river blocking, directly causing casualties and property losses. The interpretability of machine learning permits us to quantitatively assess the importance of the influencing factors, and the results of the quantitative assessment of the factors show that the six factors of distance to roads, NDVI, elevation, distance to faults, average annual rainfall, and distance to rivers have the greatest influence on landslide susceptibility, while geomorphologically related factors have a lesser combined influence on landslides.

Author Contributions: Conceptualization, methodology, visualization, and writing—original draft preparation, T.Q.; validation, formal analysis, resources, data curation, project administration, writing—review and editing, T.Q. and X.M.; software and investigation, T.Q. and Y.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by Fundamental Research Funds for the Central Universities (Grant No. lzujbky-2023-pd03), the Science and Technology Major Project of Gansu Province (Grant No. 22ZD6FA051), and the Second Tibetan Plateau Scientific Expedition and Research Program (STEP) (Grant No. 2021QZKK0201).

Data Availability Statement: Relevant data can be used in ResearchGate: Landslide locations in Bailong River Basin.

Acknowledgments: The Digital Elevation Model data were provided by the Japan Aerospace Exploration Agency (JAXA).

Conflicts of Interest: We declare that we have no financial and personal relationships with other people or organizations that can inappropriately influence our work, there is no professional or other personal interest of any nature or kind in any product, service, and/or company that could be construed as influencing the position presented in, or the review of, this manuscript.

References

1. Petley, D. Global patterns of loss of life from landslides. *Geology* **2012**, *40*, 927–930. [CrossRef]
2. Stanley, T.; Kirschbaum, D.; Zhou, Y. Spatial and temporal analysis of a global landslide catalog. *Geomorphology* **2015**, *249*, 4–15.
3. Huang, R.; Fan, X. The landslide story. *Nat. Geosci.* **2013**, *6*, 325–326. [CrossRef]
4. Shou, K.J.; Yang, C.M. Predictive analysis of landslide susceptibility under climate change conditions—A study on the Chingshui River Watershed of Taiwan. *Eng. Geol.* **2015**, *192*, 46–62. [CrossRef]
5. Chen, M.; Tang, C.; Xiong, J.; Chang, M.; Li, N. Spatio-temporal mapping and long-term evolution of debris flow activity after a high magnitude earthquake. *Catena* **2024**, *236*, 107716. [CrossRef]
6. Ayalew, L.; Yamagishi, H. The application of GIS-based logistic regression for landslide susceptibility mapping in the Kakuda-Yahiko Mountains, Central Japan. *Geomorphology* **2005**, *65*, 15–31. [CrossRef]
7. Dai, J.; Yang, J.; Yao, C.; Hu, Y.; Zhang, X.; Jiang, Q.; Zhou, C. Study on the mechanism of displacement mutation for jointed rock slopes during blasting excavation. *Int. J. Rock. Mech. Min. Sci.* **2022**, *150*, 105032. [CrossRef]
8. Peethambaran, B.; Leshchinsky, B. Application of landslide susceptibility towards urbanization suitability zonation in mountainous settings. *Int. J. Disaster Risk Reduct.* **2023**, *97*, 104061. [CrossRef]
9. Goyes-Penafiel, P.; Hernandez-Rojas, H.R. Landslide susceptibility index based on the integration of logistic regression and weights of evidence: A case study in Popayan, Colombia. *Eng. Geol.* **2021**, *280*, 105958. [CrossRef]
10. Süzen, M.L.; Doyuran, V. Data driven bivariate landslide susceptibility assessment using geographical information systems: A method and application to Asarsuyu catchment, Turkey. *Eng. Geol.* **2004**, *71*, 303–321. [CrossRef]

11. Liu, C.; Li, W.; Wu, H.; Ping, L.; Kai, S.; Sun, W.; Wen, C. Susceptibility evaluation and mapping of China's landslides based on multi-source data. *Nat. Hazards* **2013**, *69*, 1477–1495. [CrossRef]
12. Qi, T.; Zhao, Y.; Meng, X.; Chen, G.; Dijkstra, T. AI-Based Susceptibility Analysis of Shallow Landslides Induced by Heavy Rainfall in Tianshui, China. *Remote Sens.* **2021**, *13*, 1819. [CrossRef]
13. Chen, L.; Ma, P.; Yu, C.; Zheng, Y.; Zhu, Q.; Ding, Y. Landslide susceptibility assessment in multiple urban slope settings with a landslide inventory augmented by InSAR techniques. *Eng. Geol.* **2023**, *327*, 107342. [CrossRef]
14. Napoli, M.D.; Carotenuto, F.; Cevasco, A.; Confuorto, P.; Martire, D.D.; Firpo, M.; Pepe, G.; Raso, E.; Calcaterra, D. Machine learning ensemble modelling as a tool to improve landslide susceptibility mapping reliability. *Landslides* **2020**, *17*, 1897–1914. [CrossRef]
15. Zeng, T.; Wu, L.; Hayakawa, Y.S.; Yin, K.; Gui, L.; Jin, B.; Guo, Z.; Peduto, D. Advanced integration of ensemble learning and MT-InSAR for enhanced slow-moving landslide susceptibility zoning. *Eng. Geol.* **2024**, *331*, 107436. [CrossRef]
16. Ciurleo, M.; Cascini, L.; Calvello, M. A comparison of statistical and deterministic methods for shallow landslide susceptibility zoning in clayey soils. *Eng. Geol.* **2017**, *223*, 71–81. [CrossRef]
17. Zhao, Y.; Meng, X.; Qi, T.; Qing, F.; Chen, G. AI-based identification of low-frequency debris flow catchments in the Bailong River basin, China. *Geomorphology* **2020**, *359*, 107125. [CrossRef]
18. Marjanović, M.; Kovačević, M.; Bajat, B.; Voženílek, V. Landslide susceptibility assessment using SVM machine learning algorithm. *Eng. Geol.* **2011**, *123*, 225–234. [CrossRef]
19. Merghadi, A.; Yunus, A.P.; Dou, J.; Whiteley, J.; Pham, B.T.; 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]
20. Meng, X.; Chen, G.; Guo, P.; Xiong, M.; Janusz, W. Research of landslides and debris flows in Bailong River Basin: Process and prospect. *Mar. Geol. Quat. Geol.* **2013**, *33*, 1–15. [CrossRef]
21. Qi, T.; Meng, X.; Zhao, Y.; Su, X.; Chen, G.; Zeng, R.; Zhang, Y.; Li, Y.; Yue, D. Formation and distribution of landslides controlled by thrust-strike-slip fault zones and fluvial erosion in the Western Qinling Mountains, China. *Eng. Geol.* **2023**, *47*, 283–306. [CrossRef]
22. Guo, C.; Zhang, Y.; Li, X.; Ren, S.; Yang, Z.; Wu, R.; Jin, J. Reactivation of giant Jiangdingya ancient landslide in Zhouqu County, Gansu Province, China. *Landslides* **2019**, *17*, 179–190. [CrossRef]
23. Zhang, Z. Mechanism of the 2019 Yahuokou landslide reactivation in Gansu, China and its causes. *Landslides* **2020**, *17*, 1429–1440. [CrossRef]
24. Chong, Y.; Chen, G.; Meng, X.; Bian, S.; Huang, F.; Lin, L.; Yue, D.; Zhang, Y.; Guo, F. Formation mechanism and quantitative risk analysis of the landslide-induced hazard chain by an integrated approach for emergency management: A case study in the Bailong River basin, China. *Catena* **2023**, *233*, 107522. [CrossRef]
25. Li, H.; Zhang, Y.; Dong, S.; Zhang, J.; Wang, Q. Neotectonics of the Bailongjiang and Hanan faults: New insights into late Cenozoic deformation along the eastern margin of the Tibetan Plateau. *Geol. Soc. Am. Bull.* **2020**, *132*, 1845–1862. [CrossRef]
26. 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]
27. Qi, T. Spatial Distribution and Formation Mechanism of Landslides in the Bailong River Basin. Ph.D. Thesis, Lanzhou University, Gansu, China, 2022; pp. 34–51.
28. Meentemeyer, R.K.; Moody, A. Automated mapping of conformity between topographic and geological surfaces. *Comput. Geosci.* **2000**, *26*, 815–829. [CrossRef]
29. Zhang, X.; Liu, L.; Chen, X.; Gao, Y.; Xie, S.; Mi, J. GLC_FCS30: Global land-cover product with fine classification system at 30 m using time-series Landsat imagery. *Earth Syst. Sci. Data* **2020**, *182*, 2753–2776. [CrossRef]
30. Dormann, C.F.; Elith, J.; Bacher, S.; Buchmann, C.; Carl, G.; Carré, G.; Marquéz, J.R.G.; Gruber, B.; Lafourcade, B.; Leitão, P.J.; et al. Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography* **2013**, *36*, 27–46. [CrossRef]
31. Yilmaz, I. Landslide susceptibility mapping using frequency ratio, logistic regression, artificial neural networks and their comparison: A case study from Kat landslides (Tokat—Turkey). *Comput. Geosci.* **2009**, *35*, 1125–1138. [CrossRef]
32. Xu, Y.; Ju, L.; Tong, J.; Zhou, C.M.; Yang, J.J. Machine Learning Algorithms for Predicting the Recurrence of Stage IV Colorectal Cancer After Tumor Resection. *Sci. Rep.* **2020**, *10*, 2519. [CrossRef] [PubMed]
33. Zeng, J.; Zhao, Y.; Zheng, J.; Zhang, Y.; Shi, P.; Li, Y.; Chen, G.; Meng, X.; Yue, D. Early Identification of River Blockage Disasters Caused by Debris Flows in the Bailong River Basin, China. *Remote Sens.* **2024**, *16*, 1302. [CrossRef]
34. Zhang, F.; Liu, G.; Chen, W.; Liang, S.; Chen, R.; Han, W. Human-induced landslide on a high cut slope: A case of repeated failures due to multi-excavation. *J. Rock Mech. Geotech. Eng.* **2012**, *4*, 367–374. [CrossRef]
35. Vegani, C.; Giadrossich, F.; Buckley, P.; Conedera, M.; Pividori, M.; Salbitano, F.; Rauch, H.S.; Lovreglib, R.; Schwarz, M. Root reinforcement dynamics of European coppice woodlands and their effect on shallow landslides: A review. *Earth Sci. Rev.* **2017**, *167*, 88–102. [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

AI-Driven Computer Vision Detection of Cotton in Corn Fields Using UAS Remote Sensing Data and Spot-Spray Application

Pappu Kumar Yadav [1,*], J. Alex Thomasson [2], Robert Hardin [3], Stephen W. Searcy [3], Ulisses Braga-Neto [4], Sorin C. Popescu [5], Roberto Rodriguez III [6], Daniel E. Martin [7] and Juan Enciso [3]

[1] Department of Agricultural & Biosystems Engineering, South Dakota State University, Brookings, SD 57007, USA
[2] Department of Agricultural & Biological Engineering, Mississippi State University, Starkville, MS 39762, USA
[3] Department of Biological & Agricultural Engineering, Texas A&M University, College Station, TX 77843, USA; s-searcy@tamu.edu (S.W.S.)
[4] Department of Electrical & Computer Engineering, Texas A&M University, College Station, TX 77843, USA
[5] Department of Ecology & Conservation Biology, Texas A&M University, College Station, TX 77843, USA
[6] Spatial Data Analysis and Visualization Laboratory, University of Hawaii at Hilo, Hilo, HI 96720, USA
[7] Aerial Application Technology Research, U.S.D.A. Agriculture Research Service, College Station, TX 77845, USA
* Correspondence: pappu.yadav@sdstate.edu

Abstract: To effectively combat the re-infestation of boll weevils (*Anthonomus grandis* L.) in cotton fields, it is necessary to address the detection of volunteer cotton (VC) plants (Gossypium hirsutum L.) in rotation crops such as corn (*Zea mays* L.) and sorghum (*Sorghum bicolor* L.). The current practice involves manual field scouting at the field edges, which often leads to the oversight of VC plants growing in the middle of fields alongside corn and sorghum. As these VC plants reach the pinhead squaring stage (5–6 leaves), they can become hosts for boll weevil pests. Consequently, it becomes crucial to detect, locate, and accurately spot-spray these plants with appropriate chemicals. This paper focuses on the application of YOLOv5m to detect and locate VC plants during the tasseling (VT) growth stage of cornfields. Our results demonstrate that VC plants can be detected with a mean average precision (mAP) of 79% at an Intersection over Union (IoU) of 50% and a classification accuracy of 78% on images sized 1207 × 923 pixels. The average detection inference speed is 47 frames per second (FPS) on the NVIDIA Tesla P100 GPU-16 GB and 0.4 FPS on the NVIDIA Jetson TX2 GPU, which underscores the relevance and impact of detection speed on the feasibility of real-time applications. Additionally, we show the application of a customized unmanned aircraft system (UAS) for spot-spray applications through simulation based on the developed computer vision (CV) algorithm. This UAS-based approach enables the near-real-time detection and mitigation of VC plants in corn fields, with near-real-time defined as approximately 0.02 s per frame on the NVIDIA Tesla P100 GPU and 2.5 s per frame on the NVIDIA Jetson TX2 GPU, thereby offering an efficient management solution for controlling boll weevil pests.

Keywords: boll weevil; volunteer cotton (VC); remote sensing; computer vision (CV); YOLOv5; unmanned aircraft systems (UASs); spot-spray

1. Introduction

The cotton boll weevil (*Anthonomus grandis* L.) is an insect pest that has caused more than USD 23 billion in losses to the U.S. cotton industry since migrating from Mexico in the 1890s [1]. It continues to be a matter of concern for the U.S. cotton industry, particularly in Texas, even though it has been eradicated from most of the U.S. Therefore, there is a continued need for the activities of the Texas Boll Weevil Eradication Foundation (TBWEF) as per the latest report of Sunset Advisory Commission [2]. The TBWEF has divided the state into 16 eradication zones in which the Lower Rio Grande Valley (LRGV) is still actively

functioning as it remains the region most prone to boll weevil re-infestation each year due to its tropical climatic conditions [2] and proximity to the Mexico border. The vulnerability of the LRGV is due to several factors, one of which includes its subtropical climate with hot and humid summers and mild to cool winters without freezing conditions, allowing cotton plants to grow and produce fruit year-round, providing a continuous food source for boll weevils. Additionally, tropical storms from the Gulf of Mexico commonly carry boll weevils from Tamaulipas State in Mexico to the LRGV [3]. In 2019 alone, 46,000 boll weevils were captured by the foundation, indicating the severity of the problem and a continued need for the functioning of the TBWEF. Without the eradication efforts by TBWEF, boll weevils would have caused an annual loss of USD 200 million to Texas farmers [4]. Recent reports indicate that the TBWEF's efforts have significantly reduced the number of boll weevils in Texas [5]. As per the latest weekly report published by the TBWEF, the season-long trap average of boll weevils captured has decreased by approximately 99.91% from the year 2020 to the year 2024 for the Lower Rio Grande Valley [5].

Cotton (*Gossypium hirsutum* L.) is commonly planted in rotation with crops like corn (*Zea mays* L.) and sorghum (*Sorghum bicolor* L.). In climatic areas like the LRGV, cotton seeds can survive year-round, and thus, seeds in cotton that might have fallen during the harvest in the previous year can grow among corn and sorghum plants in the present year [3,6,7]. Such plants are called volunteer cotton (VC) plants, which essentially act as weeds that can emerge in fields of grain crops like corn, sorghum, soybean, and wheat. The necessity of removing VC from various crops in the state's production regions stems from two main factors: yield loss due to competition and boll weevil control challenges. Volunteer cotton plants compete with primary crops, reducing yields, with farmers generally aiming for 80–90% control using herbicides. In South and Eastern Texas, where the Texas Boll Weevil Eradication Program (TBWEP) is active, the presence of volunteer cotton significantly complicates and increases the cost of boll weevil control as boll weevils use cotton plants as hosts. Consequently, the Texas Department of Agriculture enforces a zero-tolerance policy for hostable non-commercial cotton plants (6–8 leaves or larger) in quarantined zones to prevent infestations. To comply with these regulations and protect crop yields, farmers must achieve complete control of volunteer cotton through timely and appropriate herbicide applications.

To minimize the likelihood of boll weevil re-infestation, the TBWEF uses pheromone traps to detect boll weevils and pesticides to eliminate them. As part of boll weevil mitigation efforts in the LRGV region, fields with rotation crops are inspected for the presence of VC plants at the edges of fields on a weekly basis. When VC plants are detected, the number of pheromone traps increases. In addition to inspecting for the presence of VC plants, pheromone traps are also inspected for the presence of boll weevils. If at least one is found at the edge of a field, then the entire field is sprayed with a pesticide, commonly Malathion ($C_{10}H_{19}O_6PS_2$) ULV (FYFANON® ULV AG) at rates between 0.56 and 1.12 kg/ha [8]. Uniform spraying is utilized because VC plants growing in the middle of corn and sorghum fields remain undetected and so cannot be sprayed individually. Spraying entire fields results in both increased management costs and environmental concerns as well as destroying many beneficial insects.

Uniform spray applications can be avoided if VC plants growing in the middle of corn and sorghum fields are detected and precisely located so that spot-spray-capable unmanned aircraft systems (UAS) can be deployed. Detecting VC plants before they reach the pinhead square stage and precisely spraying them with herbicides can eliminate the plants and minimize the need for applying Malathion. However, some VC plants may survive due to herbicide tolerance or inaccuracies in detection, in which case Malathion application later in the season could be needed. To detect VC plants either early or late in the growing season, remote sensing with images collected by UASs along with computer vision (CV) algorithms using state-of-the-art convolution neural network (CNN) architectures like Mask R-CNN [9], YOLOv3 [3,10,11], YOLOv5 [12], etc. can be used. Since its release in 2021, YOLOv5 has become popular in CV applications and has been used to detect

various objects like apples [13], face masks [14], safety helmets [15], etc. Due to its higher detection accuracy and faster inference speed compared to other traditional object detection algorithms, YOLOv5 was selected for this study as the most viable model for near-real-time detection. Originally YOLOv5 was released in four different variants: YOLOv5s, YOLOv5m, YOLOv5l and YOLOv5x. However, YOLOv5m was chosen as the desirable variant based on our previous research [16].

In addition to detecting VC plants, the geographic coordinates of the detected plants are needed for precise spray application. Therefore, geotagged UAS-based imagery was used in this study. High-quality RGB (Red, Green, Blue) cameras have commonly been used for object detection with YOLOv3 and YOLOv5 [13,14,17]. In most cases, radiometric correction was not employed. Remote sensing imagery without radiometric correction is susceptible to varying environmental conditions including illumination, atmospheric light scattering, sensor noise, etc. [18]. Images that are not radiometrically corrected have digital numbers (DN) that do not represent actual surface reflectance [19–21]. Hence, radiometric correction was conducted in this study.

YOLO-series object detection algorithms generate bounding boxes (BBs) around the objects of interest present in the images. The locations of these BBs are based on their pixel-wise distances from the top-left corners of images [10,22,23]. Based on the BB coordinates, central coordinates of each BB can be determined. Pixel-wise coordinates are not useful for the path planning of spot-spray-capable UASs, so a method of converting pixel-wise BB coordinates into GPS coordinates is necessary so that they can be used for path planning for spot-spraying detected VC plants.

UAS flight times are limited by battery capacity, so an optimal flight path is a necessity to efficiently spray VC plants. Optimal path planning can be conducted based on the travelling salesman problem (TSP), in which the goal is to determine the shortest route for the UAS to spot-spray each of the detected VC plants and then return to the starting point [24,25]. Different algorithms have been tested for this, some of which include the genetic algorithm used by Moon et al. [26] and Shivgan and Dong [25], ant colony optimization (ACO) used by Dorigo et al. [27], etc. In this study, ACO was used because of its simplicity in implementation, feasibility, and faster speed to generate high-quality solutions [27]. The determined optimal flight path was tested by simulating the UAS with DroneKit-Software (3.3.0) In The Loop (SITL) (3D Robotics, Berkely, CA, USA), MAVProxy [28], and Mission Planner (Ardupilot Development Team and Community). DroneKit-SITL provides an application programming interface (API) to run Python-based applications on companion computers of UASs whereas MAVProxy and Mission Planner are ground control station (GCS) software programs to control and simulate UAS applications [29]. The communication between GCS and UASs is achieved through a binary serial telemetry protocol called MAVLink [30].

The overall goal of this research was to develop a CV algorithm for detecting VC plants in a corn field and use the detected locations for optimal spot-spray applications. The four specific objectives were (i) to develop a computer vision (CV) algorithm with YOLOv5m to detect VC plants in a corn field with radiometrically (reflectance calibrated) and gamma-corrected, relatively low-resolution (1.2 Megapixel), multispectral aerial imagery; (ii) to convert the pixel-based bounding box coordinates of detected VC plants into geographic coordinates; (iii) to use the detected geographic coordinates of VC plants to generate an optimal flight path with the ACO algorithm; and (iv) to simulate a spot-spray UAS with DroneKit-SITL and the MAVProxy and Mission Planner GCS software based on the generated optimal flight path.

2. Materials and Methods

2.1. Experiment Site

This study was conducted at a corn field (Figure 1; 96°25′45.9″W, 30°32′07.4″N) of roughly 5.9 hectares (14.6 acres) at the Texas A&M AgriLife Research farm near College Station, Texas. The majority of the soil in the experimental plot is Weswood silty clay loam,

followed by Yahola fine sandy loam and Belk clay [31]. The corn plants in the field were in the second leaf (V2) vegetative state when 90 cotton seeds each of two varieties (Phytogen 340 W3FE, CORTEVA Agriscience, Indianapolis, IN, USA and Deltapine 1646 B2XF, Bayer AG, Leverkusen, North Rhine-Westphalia, Germany) were planted at randomized locations among the corn plants to mimic the presence of VC plants in the field. Some were planted in line with corn plants while others were planted in the furrow middles.



Figure 1. Experiment field located at Texas A&M University farm near College Station, TX in Burleson County (96°25′45.9″W, 30°32′07.4″N) where cotton plants were planted in the middle of corn field to mimic the presence of volunteer cotton plants.

2.2. Image Data Acquisition

A five-band (Blue, Green, Red, Near-Infrared, RedEdge) multispectral camera—RedEdge-MX (AgEagle Aerial Systems Inc., d/b/a MicaSense, Wichita, Kansas; Figure 2)—was mounted on a customized UAS (Figure 2) for collecting aerial imagery of the corn field with cotton plants when the corn plants had reached tassel vegetation stage (VT) (Figure 3). The reason for choosing this camera instead of conventional RGB camera was because its images are geotagged and can be used for other advanced image processing techniques like radiometric correction using open-source code provided by the manufacturer; moreover, it gave the flexibility to use other bands like NIR and RedEdge when required. The central wavelength and bandwidth of each spectral band are shown in Table 1. The images were acquired at this growth stage of corn plants because we wanted to test the worst-case scenario in which corn canopies are relatively larger and taller than those of VC plants. This was shown by another study in which the lowest detection accuracy was for corn fields when they had reached the VT growth stage [7]. The UAS (Hylio AG-110; Hylio Inc., Richmond, TX, USA) was originally designed for broadcast spray applications. Data were collected on 14 May 2021 between 11:00 a.m. and 2:00 p.m. central daylight-saving time (CDT) at an altitude of nearly 4.6 m (15 feet) above ground level by using auto exposure camera setting. This resulted in an approximate ground sampling distance (GSD) of 0.34 cm/pixel. This altitude was chosen for two main reasons: one was for higher image resolution and GSD while the other was so that efficient spot-spray application could be performed from this altitude without causing too much of drift in the spray droplets. Since the customized UAS was not designed for aerial surveying, there was no software interface available to capture images based on overlap settings. Therefore, images were captured

on a timer-based setting enabling an image to be captured every second. This resulted in many unused and distorted images due to unavailability of overlap settings and vibrations caused due to the design aspect of the sprayer UAS (generally, spray-capable UASs have more vibrations as compared to surveying UASs). The UAS had a payload capacity of 10 kg/L. The UAS was flown at a speed of 2 m per second.

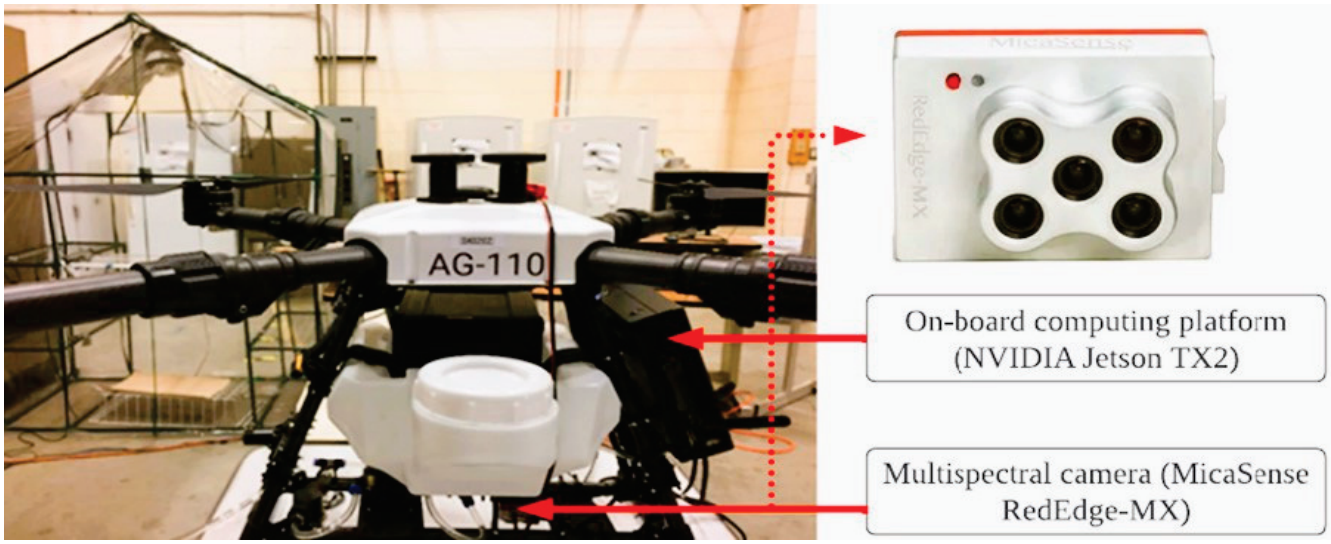


Figure 2. A customized sprayer UAS (broadcast sprayer converted to spot sprayer) with RedEdge-MX multispectral camera for capturing aerial imagery and NVIDIA Jetson TX2 computing platform [7].



Figure 3. (**A**) The customized spot-sprayer UAS flying over an experimental corn field (containing some cotton plants planted to mimic as volunteer cotton (VC) plants) capturing five band multispectral images; (**B**) RGB (Red, Green Blue) composite image showing a section of experimental plot where corn at vegetative tassel state (VT) and some cotton plants mimicking as VC plants can be seen.

Table 1. Center wavelength and full width at half-maximum (FWHM) bandwidth of each spectral band of MicaSense RedEdge-MX camera.

Spectral Band	Central Wavelength (nm)	FWHM Bandwidth (nm)
Blue	475	20
Green	560	20
Red	668	10
NIR	840	40
RedEdge	717	10

An onboard computing platform, the Jetson TX2 (NVIDIA Corporation, Santa Clara, CA, USA) development board that consists of a Pascal graphics processing unit (GPU), was mounted on the customized UAS with the intent of near-real-time detection of VC plants in corn fields (Figure 2). The Pascal GPU is low-cost, fast, and widely used as an embedded artificial intelligence computing device. It consists of 256 NVIDIA Compute Unified Device Architecture (CUDA; NVIDIA, Santa Clara, CA, USA) cores with 8 GB of RAM and 32 GB of storage capacity [32].

2.3. Manufacturer Recommended Corrections

The individual band images collected by RedEdge-MX camera were corrected based on manufacturer's recommendations [33] using the reflectance panel of type RP 04 (Figure 4) images taken on the day of data collection just before the flight.

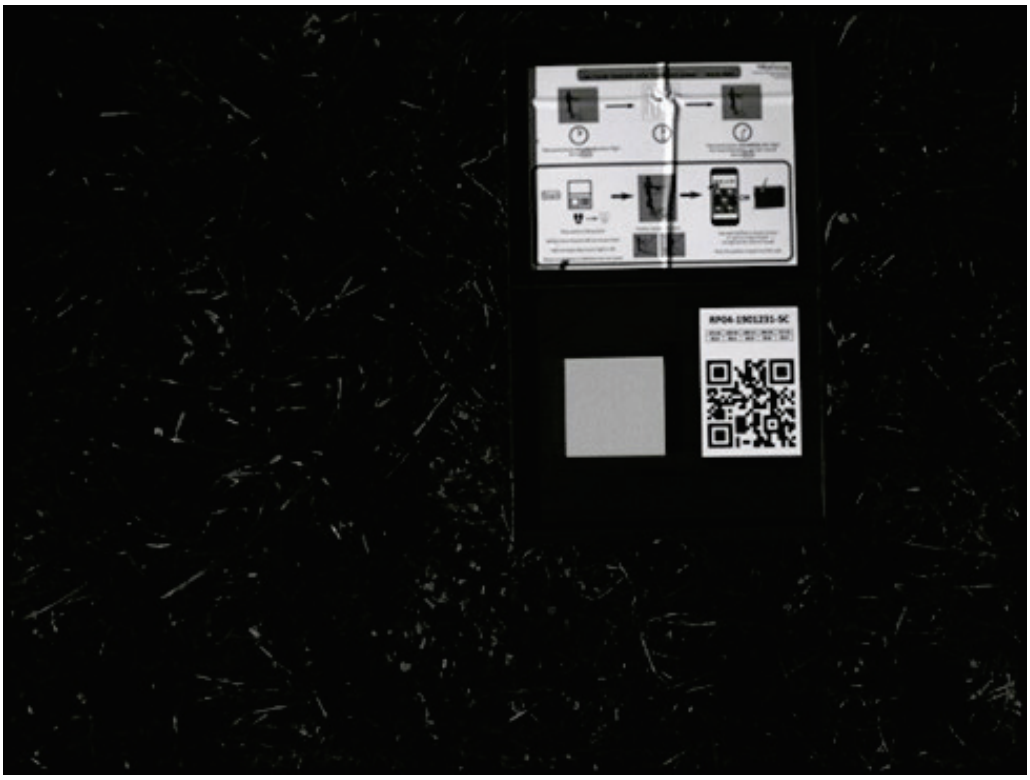


Figure 4. Reflectance panel of type RP 04 image with blue band sensor of RedEdge-MX camera taken on the day of flight.

According to the manufacturer's factory calibration (in which absolute reflectance values were plotted along wavelengths ranging from 400 nm to 850 nm), the reflectance values corresponding to blue, green, red, NIR, and RedEdge were 0.60, 0.61, 0.61, 0.60, and 0.56, respectively, for the panel that was used in the study. As per the manufacturer's process, an area of the Lambertian panel from reflectance panel image was extracted and its radiance value was converted to the scale of reflectance value, which was then applied to the whole image to convert it into reflectance images. Source codes for both radiometric and gamma corrections were used from the GitHub (GitHub, Inc., San Francisco, CA, USA) repository of MicaSense [33]. The original source code was modified based on the reflectance values for the panel used and for generating RGB aligned images for each captured image from the field data collection as seen in Figure 3. RGB images were generated because YOLOv5 network architecture accepts input images with three channels (usually RGB images). Python version 3.8.12 (Python Software Foundation, Wilmington, DE, USA) was used in Spyder integrated development environment (IDE) version 5.2.2. The entire process involves normalizing images by gain and exposure settings and then converting

them into radiance followed by reflectance. The gain and exposure settings were used to normalize the raw image data, correcting for differences in exposure time and sensor sensitivity and allowing for accurate conversion to radiance values. The normalization of images by exposure and gain settings is crucial for converting raw images to radiance. Exposure time, representing the duration the camera sensor is exposed to light, and gain, derived from the ISO setting, are obtained from the image metadata. The raw image undergoes correction to account for dark pixel offset, vignette effects, and row gradient inaccuracies. Following this, the corrected image is normalized using the exposure time and gain, ensuring consistency across different images. This normalized image data is then converted to radiance units ($W/m^2/nm/sr$) by scaling it with the gain-exposure product and applying the radiometric calibration coefficient, adjusted for the image bit depth. This process ensures accurate and comparable radiance values essential for reliable reflectance mapping. After this, in the study, unsharp mask (an image sharpening technique) was applied as enhancement technique to improve visual sharpness, and then, gamma correction was applied to make the enhanced images appear brighter and visually closer to what our eyes see [33]. For these, the original values present in the GitHub source code were used. The unsharp mask technique uses linear filter to add a fraction of high-pass-filtered input image to the original image that helps sharpen the original image by filtering out the noise [34]. Similarly, gamma correction is an image enhancement technique that is used to minimize the effect of non-linearity of the imaging sensors, thereby making the images appear brighter with enhanced contrast and visually closer [35–37].

2.4. YOLOv5

Since its release in June of 2020, YOLOv5 [38] has become a popular algorithm for object detection in CV applications. YOLOv5 was originally released in four different variants, YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x, with the subscript based on the network depth and number of parameters used. Here, *s*, *m*, *l*, and *x* represent small, medium, large, and extra-large variants of the YOLOv5 network, respectively. The YOLOv5 network comes with pretrained weights from training on the Common Objects in Context (COCO) [39] dataset, which consists of 80 different classes, such that YOLOv5 is pretrained to detect 80 different classes. In this study, the network was customized to detect a single class, "*vc*", for VC plants. The architecture of YOLOv5 comprises three main components: the Backbone, the Neck/PANet, and the Head/Output (Figure 5). The Backbone is a convolutional neural network (CNN) responsible for aggregating fine-grained images. The Neck consists of feature aggregation layers that combine features to construct the feature pyramid network. These generated feature maps are then passed on to the Head or Output network. The Output network handles the final detection phase of the model. It applies anchor boxes to the feature map obtained from the previous layer and produces a vector containing the category probability of the target object, the object score, and the position of the bounding box that surrounds the object [40].

2.5. Image Data Preparation

Among the radiometrically and gamma-corrected RGB images, only the ones containing at least a VC plant with corn in the background were chosen and then image augmentation was applied using Augmenter Python library [41]. For this, *rotate*, *flip_left_right*, *zoom_random,* and *flip_top_bottom* were used with probability values of 0.80, 0.40, 0.60, and 0.80, respectively. These values were chosen so that each time an image was passed through the augmentation pipeline, 80% of the time, *rotate* operation was applied; 40% of the time, *flip_left_right* was applied; 60% of the time, *zoom_random* was applied; and 80% of the time, *flip_top_bottom* was applied. In this way, we were able to generate a total of 521 RGB images from a total of 34 original images (containing at least one VC plant) that were radiometrically and gamma-corrected, and each of them were of the original size, 1207×923 pixels. Out of these, 417 images (80%) were used for training, 77 (15%) were

used for validation, and 27 (5%) were used for testing. The percentages of data split were chosen arbitrarily.

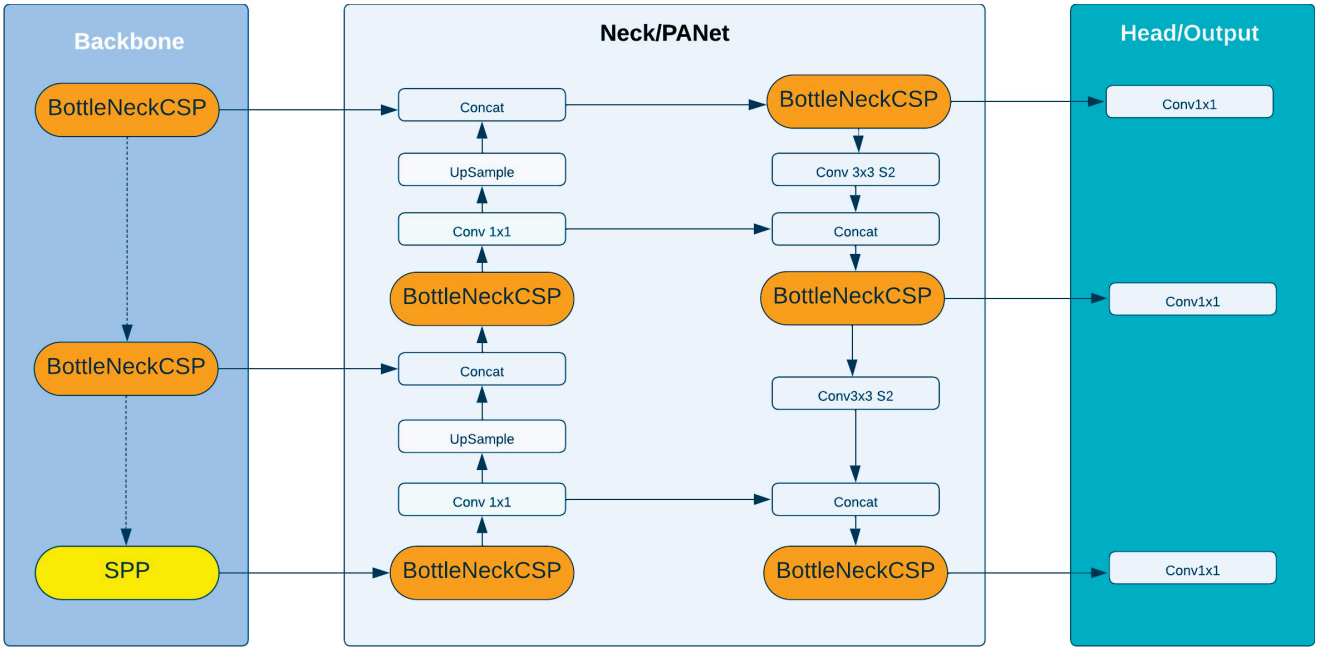


Figure 5. General overview of YOLOv5 network architecture.

2.6. YOLOv5 Training

The YOLOv5m source code was obtained from the GitHub repository of Ultralytics Inc. [12]. The PyTorch framework (Facebook AI Research Lab, Melno Park, CA, USA) with torch version 1.10.0 and Compute Unified Device Architecture (CUDA) version 11.1.0 (NVIDIA, Santa Clara, CA, USA) were used to train the YOLOv5m model on Tesla P100-PCIE-16 GB (NVIDIA, Santa Clara, CA, USA) GPU using the Google Colab (Google LLC, Melno Park, CA, USA) AI platform. The model was trained with the original hyperparameter values with initial learning rate of 0.01, final learning rate of 0.1, momentum of 0.937, weight decay of 0.0005, and intersection over union (IoU) threshold of 0.50 for a total of 621 iterations. This essentially implied that YOLOv5m had reached the convergence within the 621 iterations.

2.7. Bounding Box Coordinate Conversion

The first step involved in this process was to extract the top-left-corner coordinates from the geotagged images of RedEdge-MX camera during the processes of radiometric and gamma corrections, image enhancement, and RGB band alignments by modifying the original Python script obtained from the GitHub repository of MicaSense [33]. The extracted GPS coordinates of each geotagged image were stored in a comma-separated variable (CSV) file. The YOLOv5 *detect.py* Python script was modified to extract central coordinates of each detected bounding box, which were stored in a separate CSV file. Another Python script was developed to utilize both the CSV files and then convert the pixel-wise BB of the detected VC plants into GPS coordinates. In the Python script, GSD was first converted into meter/pixel format and the decimal coordinates were converted into Universal Transverse Mercator (UTM) format, from which northing and easting values were extracted, and then, pixel-based central coordinates were converted into UTM-based GPS coordinates using UTM zone of 14 during the conversion process [42].

Potential sources of error in this conversion process included inaccuracies in the initial geotagging of images, variations in GSD, and the precision of the UTM conversion. The accuracy of the detected bounding box coordinates was also influenced by the performance

of the YOLOv5 model, which may have had a margin of error in object detection. Additionally, any slight misalignments or distortions in the image processing steps could have propagated through to the final GPS coordinates. It was crucial to consider these factors and apply appropriate error correction and validation steps to ensure the reliability of the GPS coordinates derived from pixel-wise bounding boxes.

2.8. Optimal Flight Path with ACO Algorithm

ACO algorithm is state-of-the-art algorithm for many problems like vehicle routing, open-shop scheduling, and sequential ordering problems [27]. It is based on the fact that when ants move along a path from their colony to a food source, they deposit pheromones that evaporate over time. This means ants travelling through longer paths have less intense pheromone deposits, but the pheromone intensity is much higher along the shortest path. In our used case application, VC plant locations were the food source that the artificial ants would find through the shortest possible route. In many past cases, ACO has proven to be widely accepted algorithm for determining optimal flight paths for UASs [43,44]. Source code to implement ACO was used from the GitHub repository of *fabien-brulport* [45]. The original source code was modified to generate a CSV file containing GPS coordinates (latitude, longitude) in the order of nodes generated for the optimal route as the output from the ACO algorithm. In our study, we set the values of pheromone weight, heuristic weight, and evaporation rate as 2.01, 1, and 0.5, respectively, to achieve the shortest distance after trying with different combinations of values. Similarly, the values for number of agents, i.e., number of artificial ants (i.e., equivalence of number of UASs), and number of iterations were 1 and 100, respectively. Mathematically, at any given time t, the probability that ant k choses a path from i to j, as explained by Zhang et al. [44], is given by the following formula:

$$P_{ij}{}^{(k)}(t) = \left\{ \frac{[\tau_{ij}(t)]^{\alpha}[\eta_{ij}(t)]^{\beta}}{\sum_{S\subset \text{allowed}_k}[\tau_{is}(t)]^{\alpha}[\eta_{is}(t)]^{\beta}} \text{if } j \in \text{allowed}_k \; ; \; 0 \text{ otherwise} \right\} \quad (1)$$

Here, α and β are pheromone and heuristic weights, respectively. η_{ij} (t) and τ_{ij} (t) are visibility and quantity of pheromone between points i and j. $S \subset \text{allowed}_k$ is a set of all the possible points in a path that the ant can choose from. The pheromone amount determines the visibility, which, in turn, determines the probability of choosing a path. Therefore, higher pheromone quantity increases the visibility of a particular path between points i and j and, hence, the probability of choosing the path increases.

2.9. Spot-Spray UAS Simulation

Once the geographic coordinates were saved in optimal order corresponding to nodes obtained from ACO, the DroneKit-SITL (3D Robotics, Berkeley, CA, USA) Python package was used to simulate flight paths of UAS with a Python script and monitored on two ground control stations (GCSs): Mission Planner version 1.3.76 [46] and MAVProxy version 1.8.45 [28]. The MAVLink [30,47] protocol was used to communicate between the two GCSs and DroneKit-SITL. This communication was accomplished by using transmission control protocol (TCP) for the master port and then port-forwarding the output to three user datagram protocol (UDP) ports. The steps used to implement the entire process are represented in Algorithm 1 as shown in below.

A Python script was used to read a CSV file that contained GPS coordinates of all the nodes in order, generated by the ACO algorithm, which corresponded to the locations of the detected VC plants for spot-spraying. After the UAS simulation was performed successfully, we were also able to upload the CSV file with optimal flight paths on AgroSol version 2.105.0 GCS, developed by Hylio (Hylio, Inc., Richmond, TX, USA). The AgroSol software provided interface to control the spot-spray UAS that was customized. The flowchart in Figure 6 shows the entire workflow starting from data collection to spot-spray UAS simulation.

Algorithm 1 Spot Spray Simulation Algorithm

Require: CSV file with GPS coordinates of detected VC plants
Ensure: Simulated flight path based on ACO algorithm on Mission Planner GCS

Procedure SpotSpraySimulationAlgorithm:
Open two terminal windows from Anaconda environment;
while paths set to the directory where MAVLink was installed **do**
    if terminal window 1 **then**
    dronekit-sitl copter --home=30.534351,-96.431239,0,180
    --model=copter;
    end if
    mavproxy --master tcp:127.0.0.1:5760 --sitl 127.0.0.1:5501
    --out udp:127.0.0.1:14550 --out udp:127.0.0.1:14551 --out
udp:127.0.0.1:14552;
end while
Open Mission Planner GCS;
while mode == guided do **do**
  if UDP port == 14551 **then**
    Open Spyder IDE and execute python script "vc spot spray.py";
  else
    end
  end if
end while
end

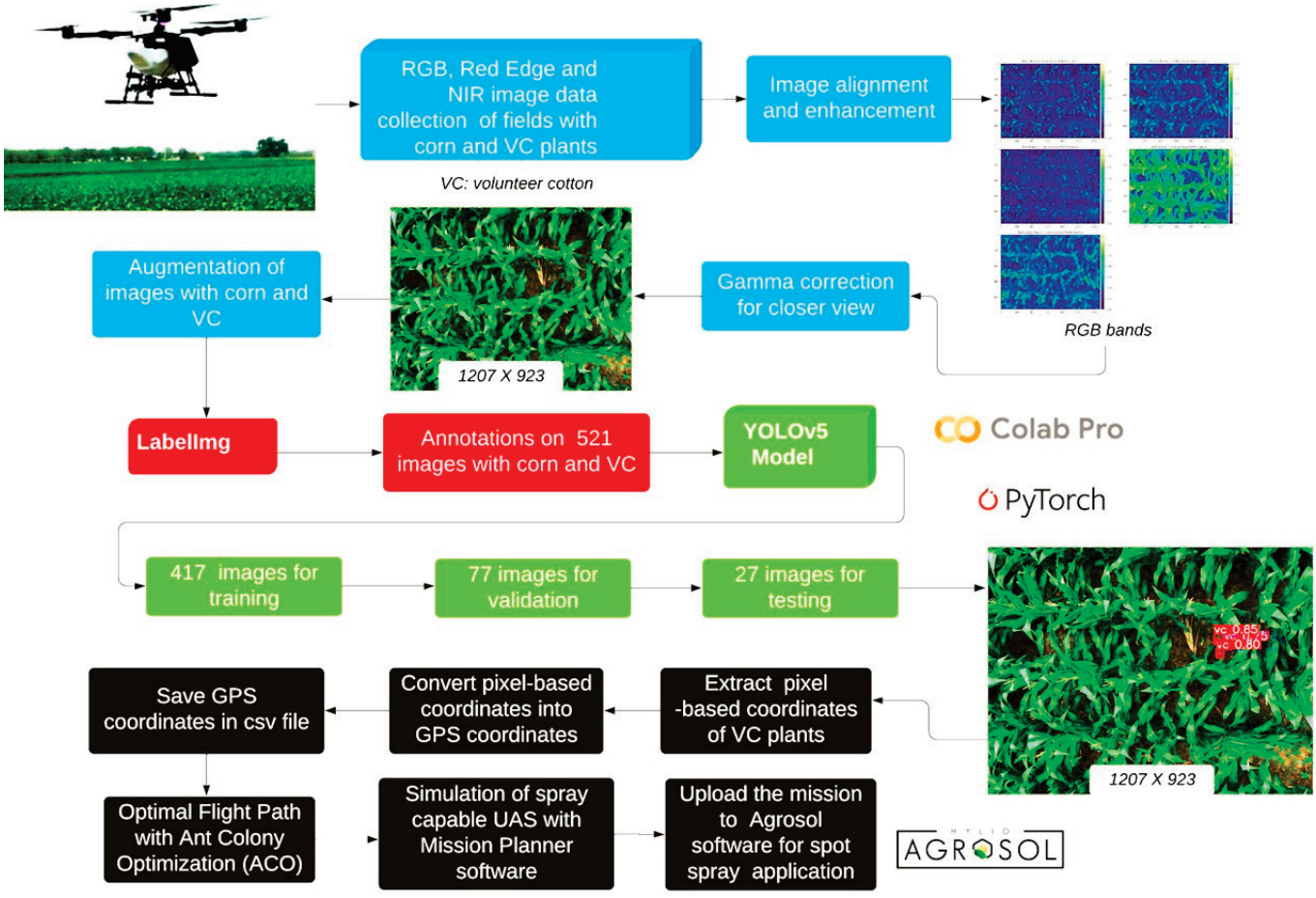


Figure 6. A flowchart that shows complete workflow representing each step used in this study.

3. Results

3.1. CV Algorithm with YOLOv5m for Detecting VC Plants in a Corn Field on Radiometrically Corrected Aerial Imagery

Figure 7 shows graphs for box loss (IoU loss) and objectness loss on both training and validation datasets. Before the 600th iteration, both the losses had converged, and no improvement was seen beyond. The lowest value of box loss was found to be 0.028 at the 476th iteration while the lowest value for objectness loss, i.e., 0.009, was found at iteration number 613 on the training dataset. On the validation dataset, the lowest value of objectness loss was found to be 0.0094 at the 394th iteration. Similarly, the lowest value of box loss on the validation dataset was found to be 0.0296 at iteration number 370. In Figure 8, graphs of different performance metrices can be seen. It was found that the maximum value of precision reached around 0.98 at the 320th iteration while the maximum value for recall was found to be 0.77 at the 525th iteration. The most important metric out of all of these was mAP@0.50, whose maximum value, i.e., 0.81, was reached at iteration number 613. The mAP@0.50 value of 0.81 indicates that the model performs well when considering detections with an Intersection over Union (IoU) threshold of 50%, meaning the predicted bounding boxes overlap with the ground truth by at least 50%. This relatively high value suggests that the model can accurately detect and localize volunteer cotton (VC) plants under these conditions. At iteration number 521, the maximum value for mAP@0.50:0.95 was obtained, and this was found to be 0.33. The mAP@0.50:0.95 value of 0.33 highlights a more nuanced picture of the model's performance across a range of IoU thresholds (from 50% to 95%). This lower value suggests that while the model is effective at identifying and localizing objects at a 50% overlap, its performance decreases as the required overlap for a correct detection becomes more challenging. This indicates potential challenges in achieving precise localization, particularly in distinguishing VC plants from the background and other objects like corn plants, weeds, and soil. The precision-recall curve (PRC) resulted in an overall accuracy of nearly 79% and the maximum value of F1-score was found to be 0.76 at nearly the 40% confidence level (Figure 9A,B). The confusion matrix in Figure 10 shows that YOLOv5m was trained enough to classify VC plants with an accuracy of 78% and loss of 22% owing to the background class of corn plants, weeds, soil, etc. To mitigate these misclassifications, several strategies can be implemented. First, increasing the diversity and size of the training dataset can help the model better distinguish between VC plants and other similar objects. This can be achieved by collecting more annotated images under various lighting conditions and growth stages. Second, data augmentation techniques such as rotation, scaling, and color adjustments can improve the model's robustness to variations in the field environment.

Figure 10 shows some detection results from the trained YOLOv5m model that resulted in an average inference speed of 47 frames per second (FPS) on NVIDIA Tesla P100 GPU-16GB. This model was later deployed on the NVIDIA Jetson TX2 GPU that was mounted on the custom UAS (Figure 2) and resulted in an adjusted average inference speed of 2.535 s (~0.4 FPS) for images of size 640 × 640 pixels (Figure 11). The inference speed of the YOLOv5m model on the NVIDIA Jetson TX2 GPU is a critical factor for assessing the feasibility of real-time applications. Real-time processing typically requires a minimum speed of 30 FPS to ensure smooth and responsive performance in dynamic environments. The observed speed of 0.4 FPS indicates a significant lag, making it unsuitable for scenarios where immediate detection and response are necessary such as in real-time monitoring and intervention in agricultural fields. This low inference speed can be attributed to multiple reasons. One possible reason could be the settings of the GPU hardware on the Jetson TX2 board, which could not utilize its maximum computing power.

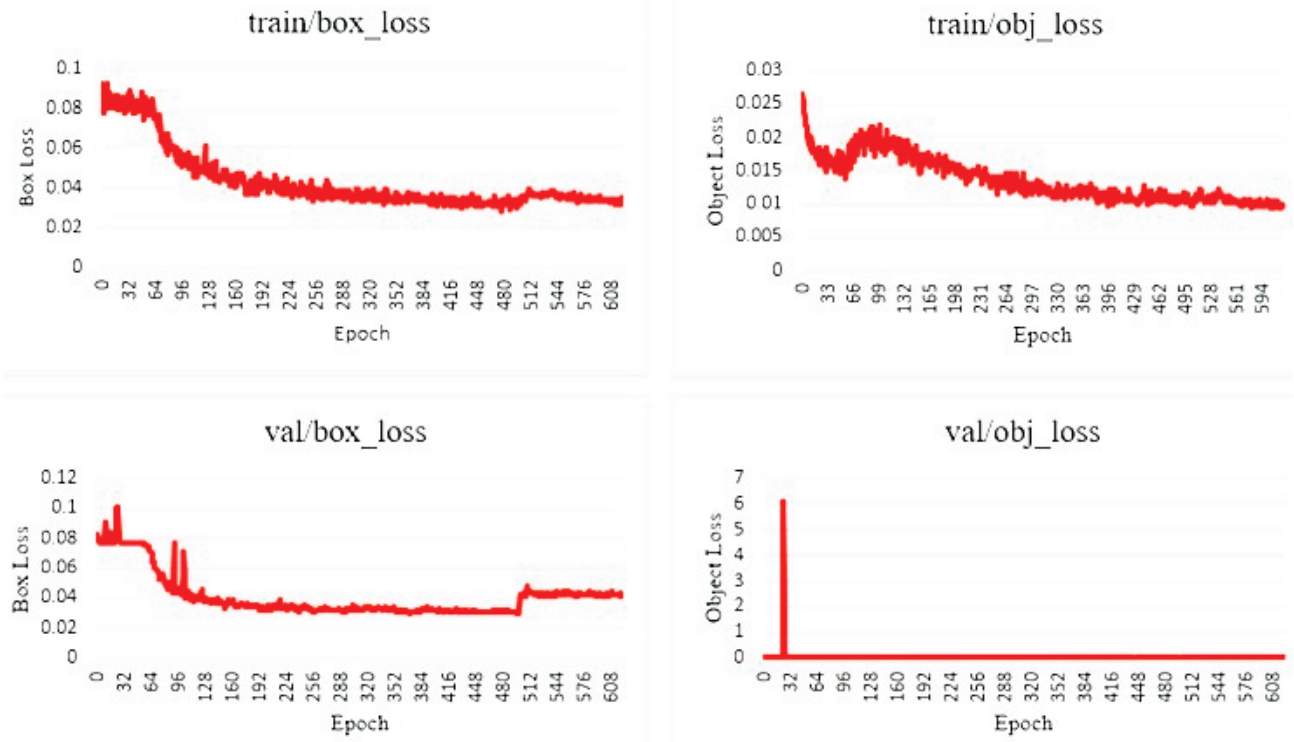


Figure 7. Different types of losses that were obtained during the training process of YOLOv5m on training and validation datasets.

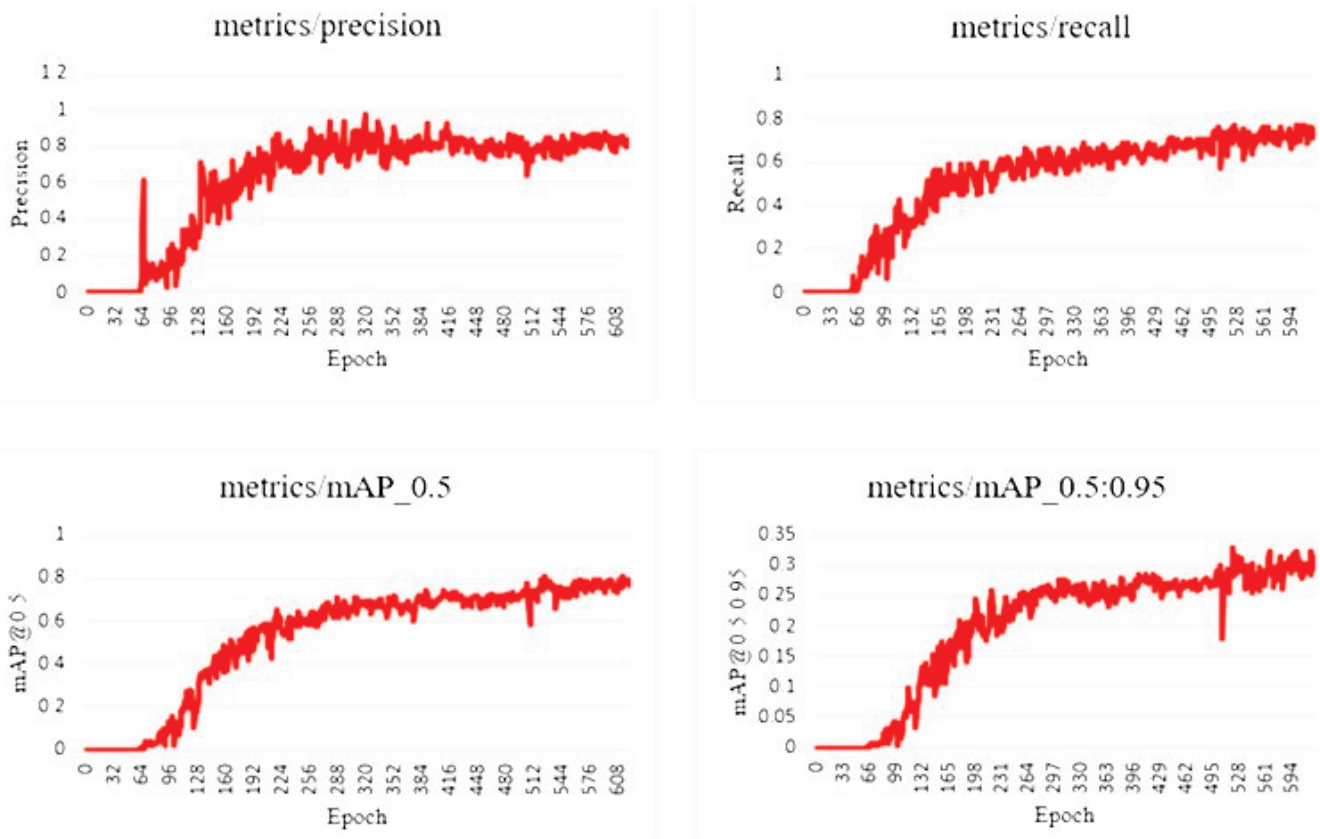


Figure 8. Different types of performance metrices that were obtained during the training process of YOLOv5m.

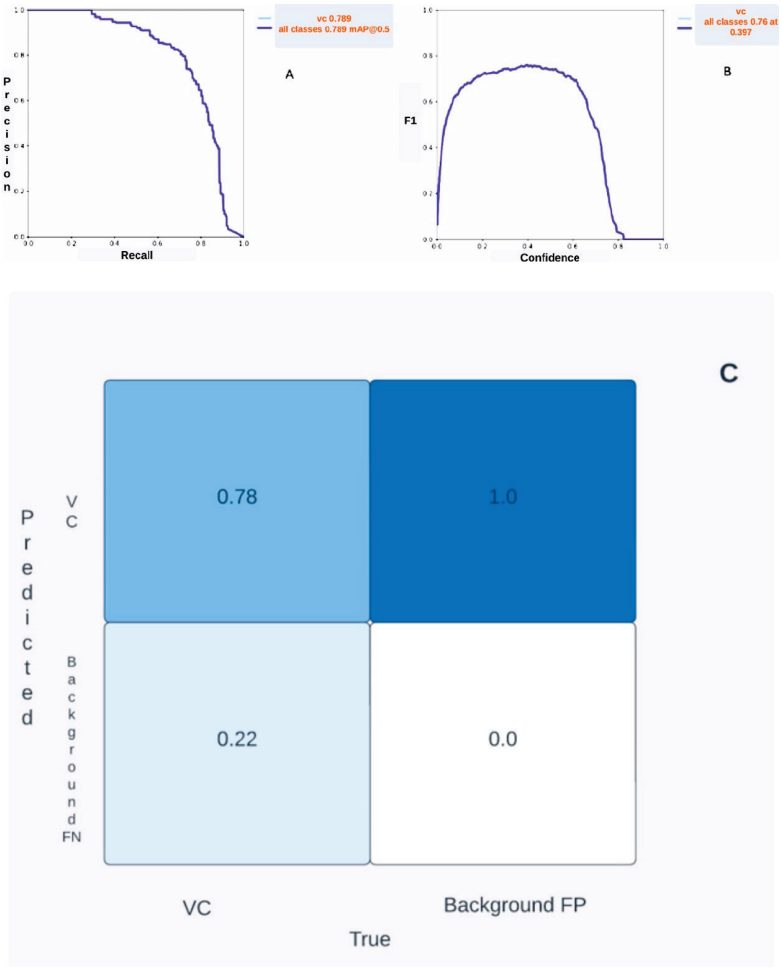


Figure 9. (**A**) Precision-recall plot, (**B**) F1-score vs confidence score plot, and (**C**) confusion matrix obtained after training YOLOv5m.

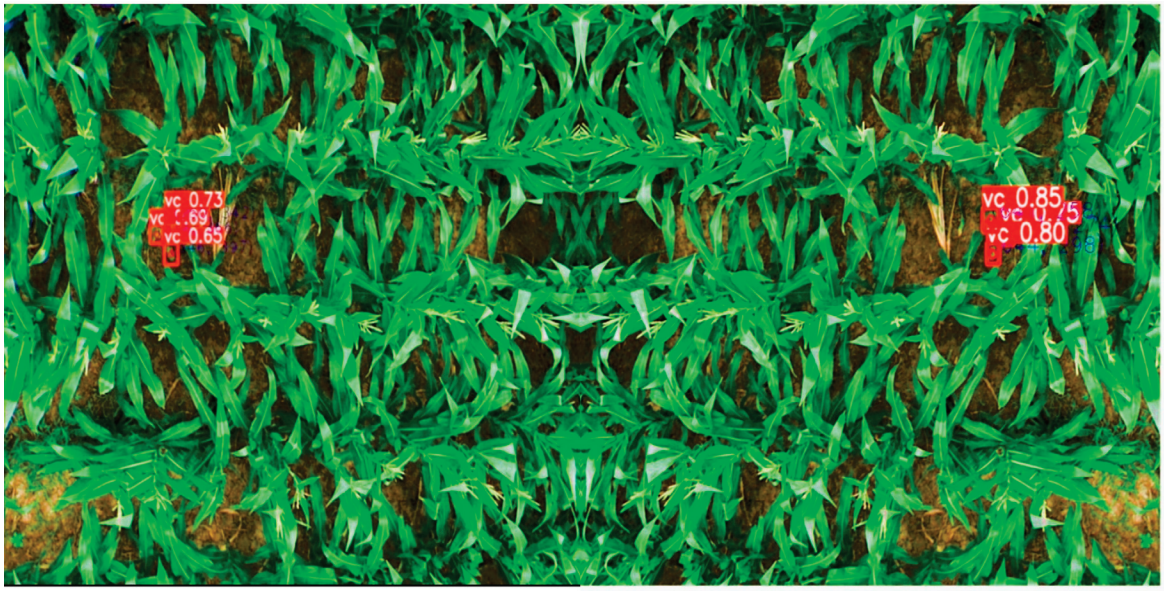


Figure 10. VC plants detected in the middle of a corn field within the red bounding boxes (BBs) by trained YOLOv5m model. The values associated with each BB show model's certainty that the bounding box contains an object of interest, i.e., VC plant.

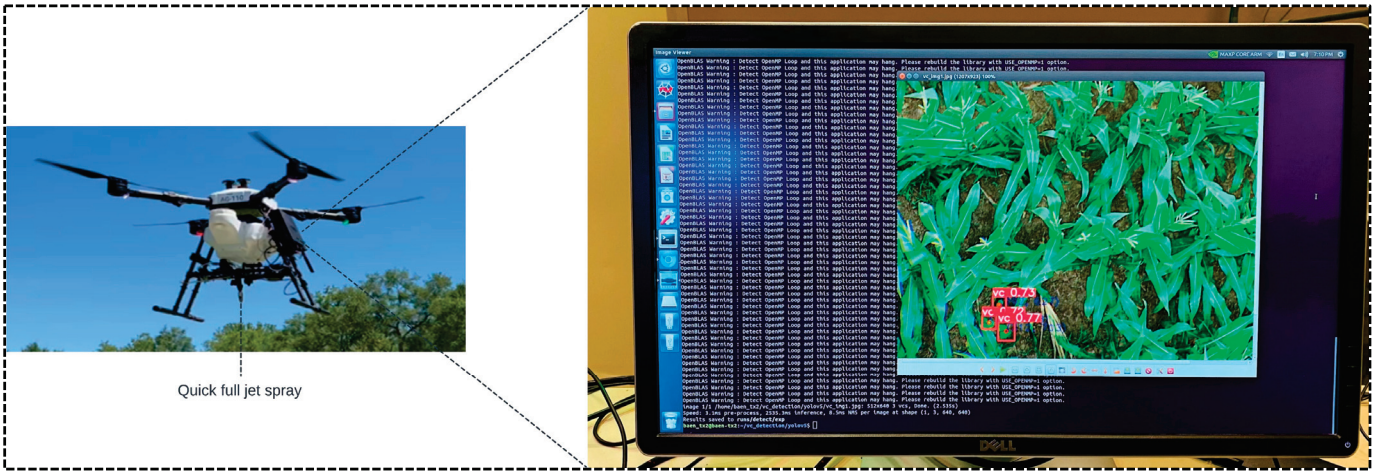


Figure 11. YOLOv5m detection of VC plants in a corn field by being deployed on NVIDIA Jetson TX2 mounted on a custom spot-spray-capable UAS.

3.2. ACO Algorithm to Determine Optimal Flight Path for Spot-Spray Applications

Ten random locations of VC plants in the experimental plot were chosen whose corresponding GPS coordinates can be seen on the left side of Table 2. The selection of ten random locations for volunteer cotton (VC) plants in the experimental plot was guided by the need to simulate realistic field conditions while ensuring the feasibility of spot-spraying in a single flight of the unmanned aerial system (UAS). These locations were chosen arbitrarily to cover a diverse range of the plot, ensuring that the UAS could efficiently manage the entire area in one flight. The criterion for selecting these locations involved ensuring that all chosen spots were accessible and could be covered within the operational constraints of the UAS such as its flight time and battery life. Each location consisted of multiple VC plants (since, at each location, multiple cotton seeds were planted); however, only one plant's GPS location was considered. This represented mimicking a real-world scenario in which VC plants usually grow in groups and where, by considering the location of one plant from each group, the entire spot consisting of the group of VC plants can be sprayed.

Table 2. GPS coordinates of randomly chosen locations of ten VC plants in experimental plot (left) and the ordered coordinates/nodes after the application of ACO algorithm (right).

VC	Latitudes	Longitudes		Nodes	Latitudes	Longitudes
1.	30.5343	−96.4312		1.	30.53687	−96.4284
2.	30.53431	−96.4301		2.	30.5359	−96.4283
3.	30.53492	−96.4303		3.	30.53321	−96.4298
4.	30.53386	−96.4299		4.	30.53386	−96.4299
5.	30.53531	−96.4289		5.	30.53431	−96.4301
6.	30.53484	−96.4299		6.	30.53492	−96.4303
7.	30.53687	−96.4284	ACO	7.	30.5343	−96.4312
8.	30.53537	−96.4289		8.	30.53484	−96.4299
9.	30.5359	−96.4283		9.	30.53537	−96.4289
10.	30.53321	−96.4298		10.	30.53531	−96.4289
				11.	30.53687	−96.4284

In this case, the first location was also considered to be the home location of the spot-spray UAS at which the flight began and ended. The right side of Table 2 shows the generated nodes in the order determined by the ACO algorithm. Nodes 1 and 11 represented the same location from which the flight began and at which it ended. The optimal path generated for spot-spray application can be seen on webpage-based output (Figure 12) that was generated by the *Streamlit* Python package (Streamlit Inc., San Francisco, CA, USA). The bottom graph shows the number of iterations on the x-axis and the optimal distance covered on the y-axis (in kilometers). The total distance covered by the spot-spray-capable UAS along the generated path was found to be 674.17 m (0.67 km/0.42 m).

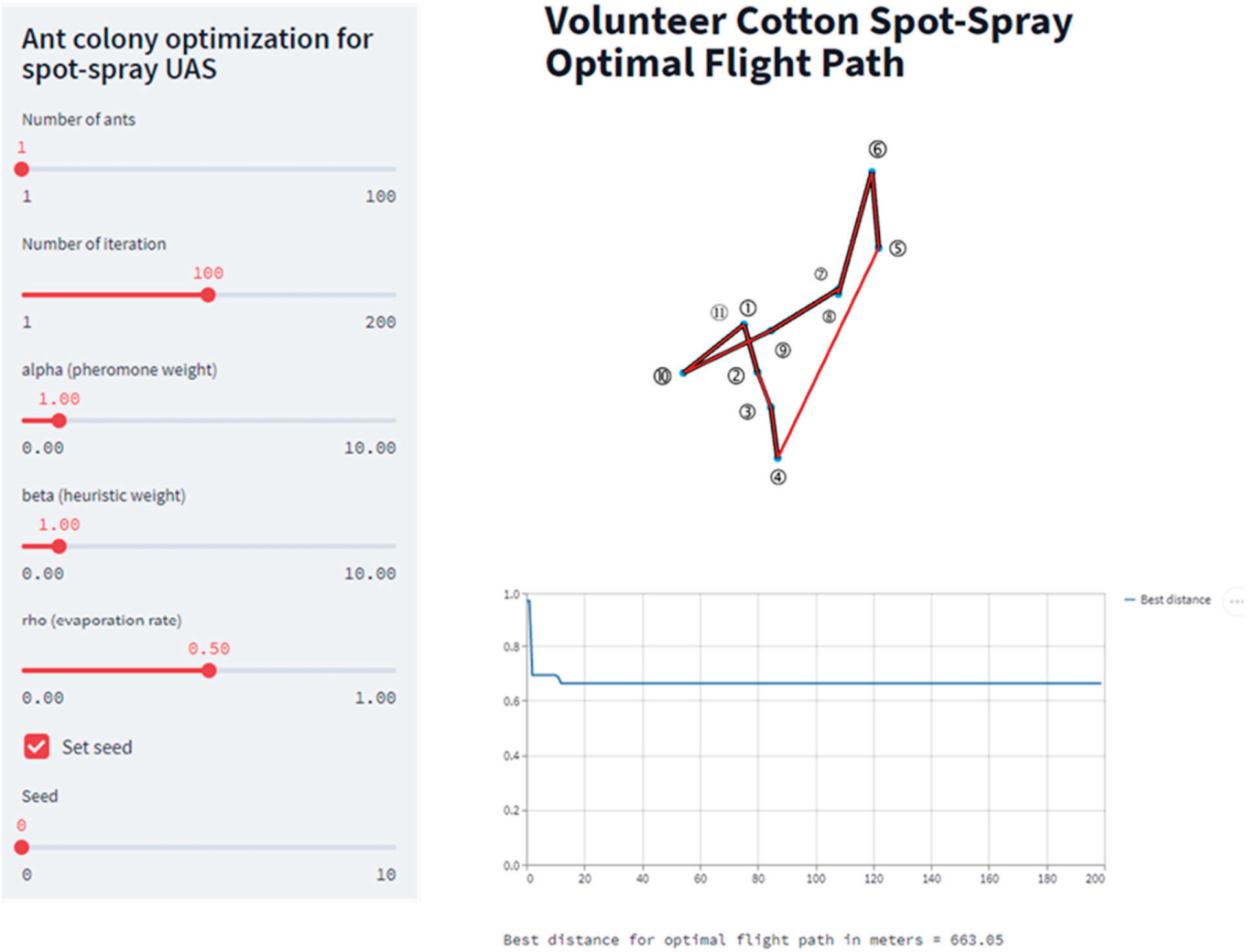


Figure 12. Optimal flight path generated by ACO algorithms and output shown by *Streamlit* Python package on a webpage.

3.3. Spot-Spray UAS Simulation on Mavproxy and Mission Planner Based on the Optimal Flight Path Generated by ACO Algorithm

Once the optimal flight path was obtained (Table 2) by the ACO algorithm, the DroneKit-SITL was used to simulate the spot-spray UAS by following the steps described under the section on *spot-spray UAS simulation* on both GCS-MAVProxy and Mission Planner as shown in Figure 13A–C.

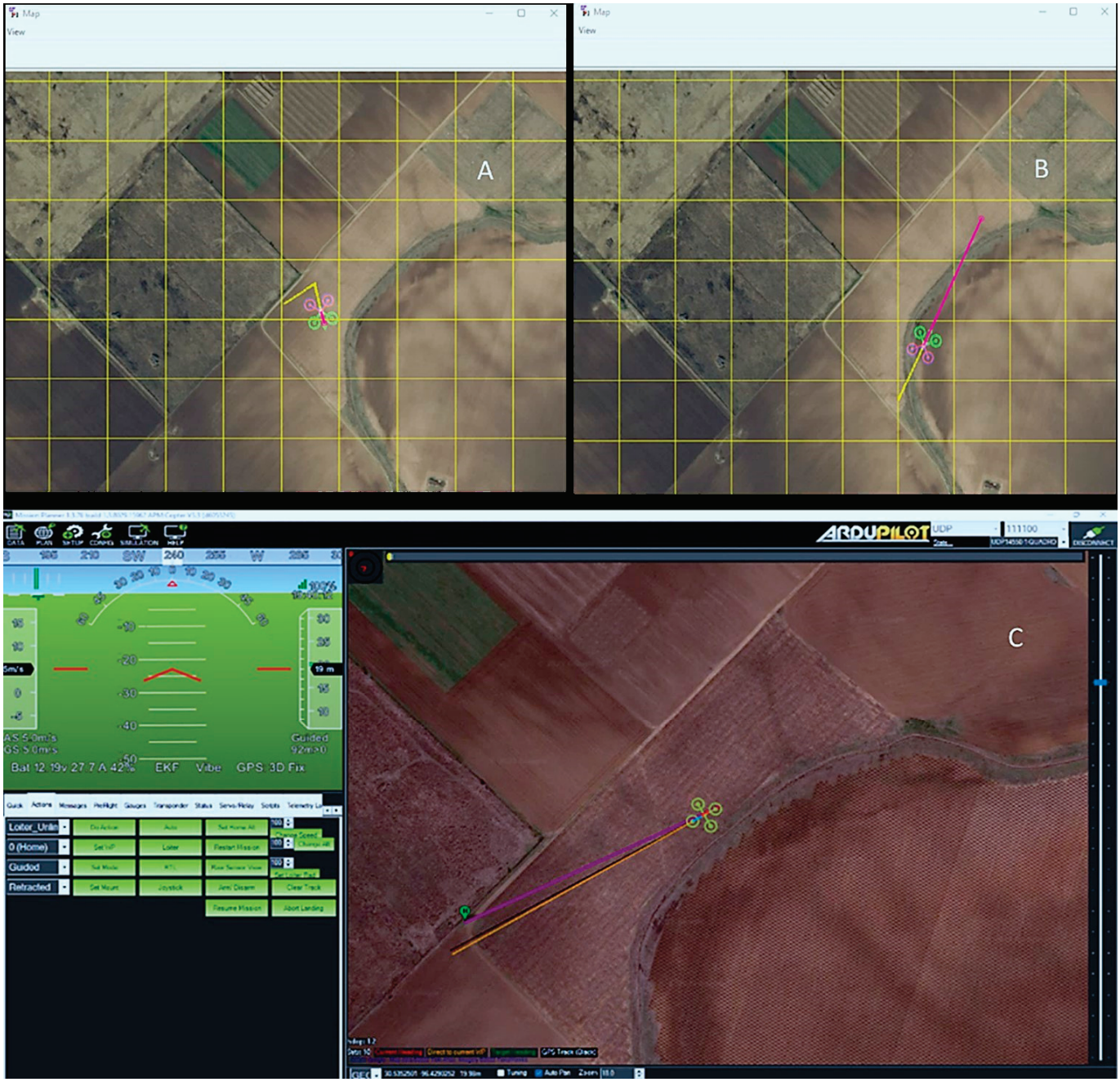


Figure 13. Spot-spray UAS simulation on MAVProxy (**A**,**B**) and Mission Planner (**C**) GCS. Image *A* shows the simulated UAS flying from node 1 to 2 while image *B* shows it flying from node 4 to 5. Image *C* shows the simulated UAS flying from node 8 to 9.

3.4. Spot-Spray Mission on Agrosol GCS

The AgroSol (Hylio Inc., Richmond, TX, USA) GCS software allows one to upload the CSV file containing GPS coordinates of nodes generated by the ACO algorithm, and then, different settings for spot-spray applications can be used for real-life spray applications (Figure 14). Once the CSV file containing the spot locations was uploaded in this study, AgroSol generated a spot group as seen on the left side of Figure 14. It then allowed us to fill in the values for different parameters like the spray altitude, spray volume, etc. The spray mission could be uploaded to the UAS flight controller, and then, the UAS could go to each of the spot locations and precisely spot-spray at each of them.

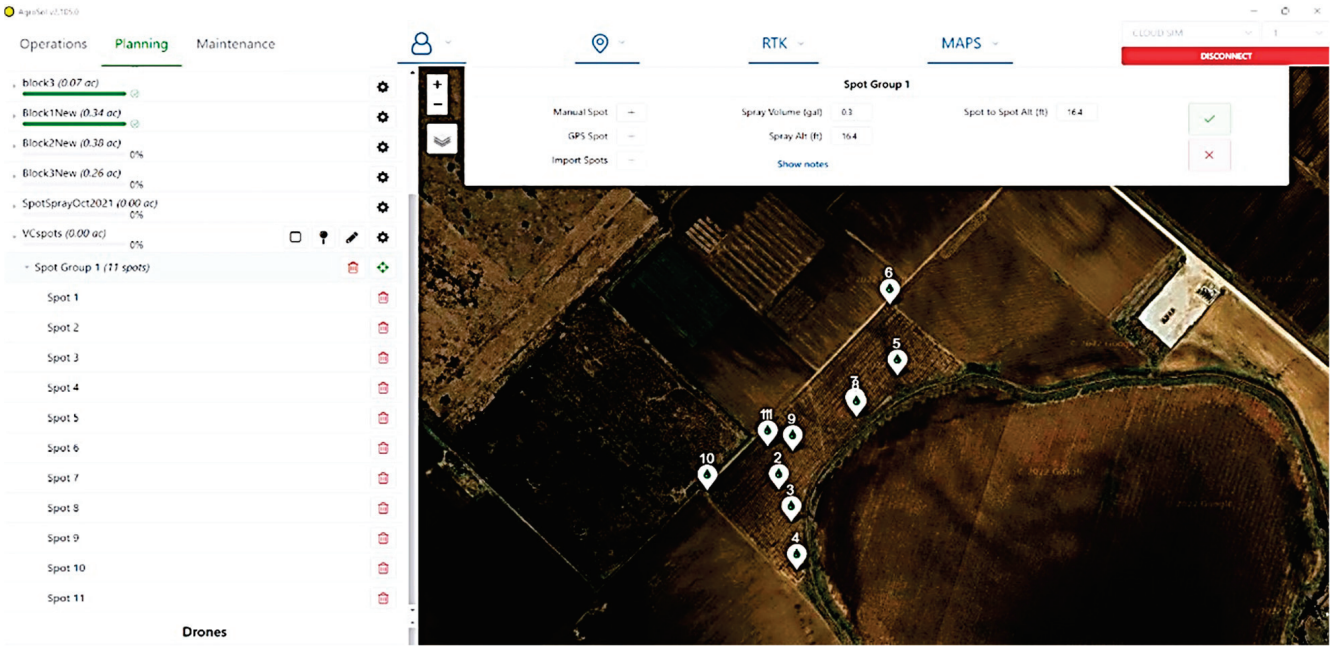


Figure 14. Spot-spray nodes generated by Agrosol software (2.87.5) after uploading the CSV file containing nodes generated by ACO algorithm.

4. Discussion

4.1. VC Detection with YOLOv5m

The potential application of multispectral remote sensing imagery for VC detection in other crops was mentioned by Zhang et al. [48] when they were developing methods of discriminating cotton plants from other crops based on spectral reflectance properties. Similarly, in our previous study [49], we were able to show that UAS-based multispectral remote sensing imagery can be used to detect VC plants growing in corn fields. In another study by Westbrook et al. [50], the authors were able to detect early-growth-stage VC plants with aerial remote sensing RGB imagery. The study reported in this paper took motivation and recommendations from all the previous studies as it made use of multispectral remote sensing imagery collected by a UAS. However, the difference lay in the use of a CV algorithm as opposed to conventional image processing techniques like discriminant and principal component analysis (PCA), maximum likelihood classification, linear spectral unmixing, etc. In the study conducted by Zhang et al. [48], they were able to discriminate cotton plants from other crops like corn, soybean, and sorghum with 100% accuracy; however, the method used a handheld spectroradiometer or the ones mounted on a tractor very close to crop canopies. This means that their method was limited by two major factors: the first one was that it could not be applied to larger areas of field and the second one was that it was a time-consuming process that was not suitable for the near-real-time detection of VC plants. Our previous study [49] used remote sensing multispectral imagery and some classical machine learning techniques to make the process of VC detection semi-automatic; it, however, could not produce a classification accuracy greater than 70%.

In this study, we were able to develop a CV algorithm with YOLOv5m to detect VC plants (before the pinhead square growth stage) growing in a corn field at the tassel (VT) growth stage using multispectral remote sensing imagery. In another study [3], we were able to detect VC plants with more than 90% accuracy in an early-growth corn field (V3 vegetative growth stage) with uncalibrated RGB imagery. YOLOv3 was used, which resulted in a nearly 29% higher classification accuracy than the previous approach [49], but the trained model was less robust to illumination conditions and environmental factors as the images were not radiometrically corrected. Moreover, the relative height difference between VC and corn plants was less compared to that in the study reported in this paper.

In our current and latest approach, reported herein, we were able to develop a CV algorithm based on YOLOv5m, which was more robust to illumination conditions and environmental factors as it was trained on radiometrically corrected imagery [51]. By training the YOLOv5m model on radiometrically corrected imagery, we were able to generate a more reliable model because it was trained on more accurate data [21]. With the current approach, we were able to classify VC plants at 11.43% higher accuracy than the previous method [49] as seen in Figure 10. It is also noteworthy that similar detection accuracy was obtained in a corn field when the plants were at the VT stage unlike the V3 stage reported in the other study [3]. YOLOv5 offers several advantages over its predecessors and other detection models such as Faster R-CNN, EfficientDet, and SSD (Single Shot MultiBox Detector) that have their own strengths and weaknesses. Faster R-CNN is known for its high accuracy, but it often comes at the cost of slower inference speeds, making it less suitable for real-time applications. EfficientDet balances accuracy and speed effectively, but it requires extensive hyperparameter tuning and larger computational resources. SSD offers good speed but generally lags in accuracy compared to YOLOv5.

The overarching goal of this study was to speed up the management aspect of the Texas Boll Weevil Eradication Program (TBWEP) by simultaneously reducing the chemical costs. Therefore, apart from developing a system for the near-real-time detection of VC plants, it is also required to develop a detection algorithm that results in minimal false negatives (i.e., maximum recall). This way, we can minimize the possibility of missing VC plants in the middle of corn fields that can potentially act as hosts for boll weevil pests. In other words, for our used case scenario, our algorithm can be tolerant to false positives but not to false negatives as the damages incurred by boll weevil pests are much more than the extra costs incurred due to spraying some undesired locations because of false positives. For this reason, the model obtained at the 525th iteration can be used (Figure 8). However, if high precision is required (like in the case of saving chemical costs), then the model trained at iteration 322 can be used, resulting in a precision as high as 98% (Figure 8). To quantitatively analyze the trade-offs between precision and recall, we examined the performance of the YOLOv5m model at different iterations. The model at the 525th iteration demonstrated a maximum recall of 0.77, which was critical for minimizing the possibility of missing any VC plants. This high recall value indicates that the model could detect most of the VC plants, thus preventing potential boll weevil infestations. However, this came at the cost of a lower precision, meaning there were more false positives. In our context, this trade-off is acceptable as the primary objective is to ensure no VC plants are missed. The overall performance of the model is represented by mAP@50, whose maximum value was found to be 81% at the 613th iteration (Figure 8). The performance of the trained model can be improved by adjusting the confidence threshold value mentioned by Yan et al. [40]. Apart from these, we had generated 521 images from a limited number of datasets, i.e., 34 images, by using image augmentation techniques. In this way, the trained model reported in this study can be regarded as relatively more generalizable [52]. While image augmentation techniques can enhance the diversity of a training set, they may not fully capture the variability present in real-world scenarios such as that due to different lighting conditions, plant growth stages, and environmental backgrounds. To improve the generalizability of the model, several strategies can be implemented, one of which is increasing the size and diversity of the training dataset by collecting more annotated images from various field conditions and growth stages, which can help the model learn more robust and generalizable features.

4.2. ACO Algorithm for Optimal Flight Path and Spot-Spray

UAS flight parameters such as aerial speed, flight altitude, yaw angle, etc. were assumed to be fixed and the optimal path generated, shown in Figure 12, was based on equal weights assigned to both pheromones and heuristic parameters. These assumptions were like the ones made by Ma et al. [53]. The only constraint used in our implementation was the distance between two GPS locations. However, in other studies, parameters like the

yaw angle were used as some of the constraints [44]. The ACO algorithm being stochastic in nature generates sufficiently good solutions based on randomly generated variables but no globally optimal solution [54]. Therefore, the optimal path shown in Figure 14 may not be the global optimal solution but simply one based on the randomly generated variables used in our study. The parameter values used for the ACO in this study may not be the best ones but were tested with reasonable combinations. From past studies, it was found that ACO is highly sensitive to the evaporation rate, which is considered to be the equivalent of the learning rate; therefore choosing the right value of this parameter is crucial [55,56]. Ojha et al. [55] found that the performance of ACO decreased beyond the 0.5 value for the evaporation rate. Hence, we chose to use 0.5 as the evaporation rate in our case (Figure 12). A total of 11 nodes were generated from the 10 VC plant locations, as seen in Table 2—left and 2—right. Nodes 1 and 11 represent the same location from which the UAS began and ended its journey traversing through the optimal path generated (Figures 12–14).

5. Conclusions and Future Work

5.1. Conclusions

This paper has presented our findings on the successful application of a CV algorithm based on YOLOV5m for the detection of VC plants in a corn field during the tassel growth stage using UAS remote sensing multispectral imagery. Our approach showcased improved classification and detection accuracies compared to previous methods. Additionally, we demonstrated that a low-resolution imaging sensor, coupled with preprocessing algorithms such as radiometric and gamma corrections, can effectively detect VC plants while maintaining robustness against variations in illumination and environmental conditions.

Furthermore, we showcased the practicality of the trained YOLOV5m model for near-real-time detection by deploying it on a computing platform mounted on a sprayer UAS. This implementation enabled the detection of VC plants in full-scale images (1207 $\times$ 923 pixels) at an adjusted average inference speed of approximately 0.4 FPS on a Pascal GPU of the Jetson TX2 development board from NVIDIA.

Moreover, we successfully converted the pixel-wise bounding box central coordinates of the detected VC plants into GPS coordinates. This allowed us to generate optimal flight paths using the ACO algorithm for simulating spot-spray applications. Overall, our research has led to the development of a CV algorithm capable of near-real-time VC plant detection in corn fields along with spot-spray applications utilizing a customized spray-capable UAS. Through this work, we have created a system with the potential to expedite TBWEP mitigation efforts possibly at a reduced management cost.

5.2. Limitations and Future Work

We encountered several limitations in our study. Firstly, we only had 34 original images that contained VC plants. To address this issue, we applied image augmentation techniques to the entire set of original images. Although this approach could potentially lead to data leakage, Saulo Barreto [57] has shown that it can still be used effectively in cases with limited data. Another limitation was observed in our study regarding the slow inference speed of VC detection when deploying the trained YOLOv5m model on the Jetson TX2 platform. This limitation can be mitigated by configuring the hardware appropriately and deploying a lighter version of YOLOv5 for real-time applications. Lastly, the simulation results of spot-spray applications were not validated under field conditions. Therefore, our future work will involve conducting field tests to validate the simulated results.

Author Contributions: Conceptualization, P.K.Y. and J.A.T.; methodology, P.K.Y. and J.A.T.; software, P.K.Y.; validation, P.K.Y. and J.A.T.; formal analysis, P.K.Y. and J.A.T.; investigation, P.K.Y., J.A.T., R.H., S.W.S., U.B.-N., S.C.P., R.R.III and D.E.M.; resources, J.A.T., R.H. and J.E.; data curation, P.K.Y. and J.A.T.; writing—original draft preparation, P.K.Y.; writing—review and editing, P.K.Y., J.A.T., R.H., S.W.S., U.B.-N., S.C.P., R.R.III, D.E.M. and J.E.; visualization, P.K.Y. and J.A.T.; supervision, J.A.T. and

R.H.; project administration, J.A.T. and R.H.; funding acquisition, J.A.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded and supported by the United States Department of Agriculture's Animal and Plant Health Inspection Service (APHIS) Cooperative Agreement# AP20PPQS&T00C046.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: All data that support the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgments: Sincere thanks are extended to the farm manager (Stephen P. Labar), the student assistants (Roy Graves, Madison Hodges, Sam Pyka, Reese Rusk, Raul Sebastian, JT Womack, John Marshall, Katelyn Meszaros, Lane Fisher, and Reagan Smith) involved during the field work, and all the unanimous reviewers whose feedback helped improve the quality of this paper. We would also like to thank USDA-APHIS for supporting this project.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Harden, G.H. *Texas Boll Weevil Eradication Foundation Cooperative Agreement*; United States Department of Agriculture: Washington, DC, USA, 2018.
2. Roming, R.; Leonard, A.; Seagraves, A.; Miguel, S.S.; Jones, E.; Ogle, S. Sunset Staff Reports with Final Results. 2021. Available online: www.sunset.texas.gov (accessed on 23 July 2024).
3. Yadav, P.K.; Thomasson, J.A.; Hardin, R.; Searcy, S.W.; Braga-Neto, U.; Popescu, S.C.; Martin, D.E.; Rodriguez, R.; Meza, K.; Enciso, J.; et al. Detecting volunteer cotton plants in a corn field with deep learning on UAV remote-sensing imagery. *Comput. Electron. Agric.* **2023**, *204*, 107551. [CrossRef]
4. Wayne, R. Texas Department of Agriculture Commissioner Sid Miller. Texas Department of Agriculture. Available online: https://www.texasagriculture.gov/News-Events/Article/3021/Commissioner-Miller-Announces-Successes-for-Boll-Weevil-Eradication-in-Texas (accessed on 18 June 2024).
5. Texas Boll Weevil Eradication Foundation, Inc. Weekly Report. 2024. Available online: https://www.txbollweevil.org/Zones/WeeklyMaster.pdf (accessed on 18 June 2024).
6. Wang, T.; Mei, X.; Thomasson, J.A.; Han, X.; Yadav, P.K. GIS-based volunteer cotton habitat prediction and plant-level detection with UAV remote sensing. *Comput. Electron. Agric.* **2022**, *193*, 106629. [CrossRef]
7. Yadav, P.K.; Thomasson, J.A.; Searcy, S.W.; Hardin, R.G.; Braga-Neto, U.; Popescu, S.C.; Martin, D.E.; Rodriguez, R.; Meza, K.; Enciso, J.; et al. Assessing the performance of YOLOv5 algorithm for detecting volunteer cotton plants in corn fields at three different growth stages. *Artif. Intell. Agric.* **2022**, *6*, 292–303. [CrossRef]
8. FMC Corporation. "FYFANON ULV AG," *FYFANON ULV AG*. Philadelphia. 2001. Available online: https://www.sciencedirect.com/science/article/abs/pii/B9780815513810500075?via=ihub (accessed on 18 June 2024).
9. He, K.; Gkioxari, G.; Dollar, 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. 2961–2969.
10. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767. [CrossRef]
11. Yadav, P.K.; Thomasson, J.A.; Hardin, R.G.; Searcy, S.W.; Braga-Neto, U.M.; Popescu, S.C.; Martin, D.E.; Rodriguez, R.; Meza, K.; Enciso, J.; et al. Volunteer cotton plant detection in corn field with deep learning. In *Autonomous Air and Ground Sensing Systems for Agricultural Optimization and Phenotyping VII*; SPIE: Orlando, Fl, USA, 2022; p. 3. [CrossRef]
12. Jocher, G.; Changyu, L.; Hogan, A.; Yu, L.; Rai, P.; Sullivan, T. YOLOv5, ultralytics/yolov5: Initial Release. Available online: https://github.com/ultralytics/yolov5/tree/v1.0 (accessed on 20 February 2023).
13. Kuznetsova, A.; Maleva, T.; Soloviev, V. YOLOv5 versus YOLOv3 for Apple Detection. In *Cyber-Physical Systems: Modelling and Intelligent Control*, 338th ed.; Kravets, A.G., Bolshakov, A.A., Shcherbakov, M., Eds.; Springer Nature: Warsaw, Poland, 2021.
14. Sharma, V. *Face Mask Detection using YOLOv5 for COVID-19*; California State University-San Marcos: San Marcos, CA, USA, 2020.
15. Zhou, F.; Zhao, H.; Nie, Z. Safety Helmet Detection Based on YOLOv5. In Proceedings of the 2021 IEEE International Conference on Power Electronics, Computer Applications (ICPECA), Shenyang, China, 22–24 January 2021; pp. 6–11. [CrossRef]
16. Yadav, P.K.; Thomasson, J.A.; Hardin, R.; Searcy, S.W.; Braga-Neto, U.; Popescu, S.C.; Rodriguez, R.; Martin, D.E.; Enciso, J.; Meza, K.; et al. Plastic Contaminant Detection in Aerial Imagery of Cotton Fields Using Deep Learning. *Agriculture* **2023**, *13*, 1365. [CrossRef]
17. Li, G.; Suo, R.; Zhao, G.; Gao, C.; Fu, L.; Shi, F.; Dhupia, J.; Li, R.; Cui, Y. Real-time detection of kiwifruit flower and bud simultaneously in orchard using YOLOv4 for robotic pollination. *Comput. Electron. Agric.* **2022**, *193*, 106641. [CrossRef]
18. Hausamann, D.; Zirnig, W.; Schreier, G.; Strobl, P. Monitoring of gas pipelines—A civil UAV application. *Aircr. Eng. Aerosp. Technol.* **2005**, *77*, 352–360. [CrossRef]

19. Minařík, R.; Langhammer, J.; Hanuš, J. Radiometric and atmospheric corrections of multispectral μMCA Camera for UAV spectroscopy. *Remote Sens.* **2019**, *11*, 2428. [CrossRef]
20. Biday, S.G.; Bhosle, U. Relative Radiometric Correction of Multitemporal Satellite Imagery Using Fourier and Wavelet Transform. *J. Indian Soc. Remote Sens.* **2012**, *40*, 201–213. [CrossRef]
21. Mamaghani, B.; Salvaggio, C. Multispectral sensor calibration and characterization for sUAS remote sensing. *Sensors* **2019**, *19*, 4453. [CrossRef]
22. Redmon, J.; Sinigardi, S.; Hager, T.; Maaz, M.; Zhang, V.; Alasuutari, J.; Kahn, P.; Ovodov, I.; Veitch-Michaelis, J.; Dujardin, A.; et al. AlexeyAB/darknet: YOLOv3. Available online: https://zenodo.org/record/5622675 (accessed on 6 June 2022).
23. Jocher, G.; Stoken, A.; Borovec, J.; Christopher, S.T.; Laughing, L.C. Ultralytics/yolov5: v4.0-nn.SiLU() activations, Weights & Biases logging, PyTorch Hub integration. *Zenodo* **2021**. [CrossRef]
24. Sorma, R.; Wadud, A.; Karim, S.M.R.; Ahamed, F.A.S. Solving Traveling Salesman Problem by Using Genetic Algorithm. *Electr. Electron. Eng.* **2020**, *10*, 27–31. [CrossRef]
25. Shivgan, R.; Dong, Z. Energy-Efficient Drone Coverage Path Planning using Genetic Algorithm. In Proceedings of the 2020 IEEE 21st International Conference on High Performance Switching and Routing (HPSR), Newark, NJ, USA, 11–14 May 2020. [CrossRef]
26. Moon, C.; Kim, J.; Choi, G.; Seo, Y. An efficient genetic algorithm for the traveling salesman problem with precedence constraints. *Eur. J. Oper. Res.* **2002**, *140*, 606–617. [CrossRef]
27. Dorigo, M.; Birattari, M.; Stutzle, T. Ant Colony Optimization. *Stud. Comput. Intell.* **2021**, *947*, 3–8. [CrossRef]
28. Tridgell, A.; Barker, P. ArduPilot MAVProxy. Available online: https://ardupilot.org/mavproxy/index.html (accessed on 2 March 2022).
29. Qays, H.M.; Jumaa, B.A.; Salman, A.D. Design and Implementation of Autonomous Quadcopter using SITL Simulator. *Iraqi J. Comput. Commun. Control. Syst. Eng.* **2020**, 1–16. [CrossRef]
30. Meier, L. MAVLink Developer Guide. Available online: https://mavlink.io/en/ (accessed on 19 May 2021).
31. USDA-Natural Resources Conservation Service, "Web Soil Survey". Available online: https://websoilsurvey.sc.egov.usda.gov/App/HomePage.htm (accessed on 19 July 2021).
32. NVIDIA. Jetson TX2 Module. Available online: https://developer.nvidia.com/embedded/jetson-tx2 (accessed on 19 September 2021).
33. MicaSense Incorporated. MicaSense RedEdge and Altum Image Processing Tutorials. Available online: https://github.com/micasense/imageprocessing (accessed on 14 October 2021).
34. Allebach, J.P. Optimal unsharp mask for image sharpening and noise removal. *J. Electron. Imaging* **2005**, *14*, 023005. [CrossRef]
35. Guo, H.; He, H.; Chen, M. Gamma correction for digital fringe projection profilometry. *Appl. Opt.* **2004**, *43*, 2906–2914. [CrossRef] [PubMed]
36. Ju, M.; Ding, C.; Zhang, D.; Guo, Y.J. Gamma-correction-based visibility restoration for single hazy images. *IEEE Signal Process. Lett.* **2018**, *25*, 1084–1088. [CrossRef]
37. Xu, G.; Su, J.; Pan, H.; Zhang, Z.; Gong, H. An image enhancement method based on gamma correction. In Proceedings of the 2009 Second International Symposium on Computational Intelligence and Design, Changsha, China, 12–14 December 2009; pp. 60–63. [CrossRef]
38. Jocher, G.; Stoken, A.; Borovec, J.; Chaurasia, A.; Liu, C.; Hajek, J.; Diaconu, L.; Defretin, Y.; Lohia, A.; Milanko, B.; et al. ultralytics/yolov5: v5.0-YOLOv5-P6 1280 Models, AWS, Supervise.ly and YouTube Integrations. Available online: https://zenodo.org/record/4679653 (accessed on 21 July 2021).
39. 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 *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2014; pp. 740–755. [CrossRef]
40. Yan, B.; Fan, P.; Lei, X.; Liu, Z.; Yang, F. A real-time apple targets detection method for picking robot based on improved YOLOv5. *Remote. Sens.* **2021**, *13*, 1619. [CrossRef]
41. Bloice, M.D. Augmentor: Image Augmentation Library in Python for Machine Learning. August 2017. Available online: https://zenodo.org/records/1041946 (accessed on 5 August 2022).
42. "State of Texas UTM Zones," Texas Parks & Wildlife. Available online: https://tpwd.texas.gov/publications/pwdpubs/media/pwd_mp_e0100_1070ah_08.pdf (accessed on 5 November 2021).
43. Sun, Y.; Chen, J.; Du, C. Path planning of UAVs based on improved ant colony system. In Proceedings of the 2020 IEEE International Conference on Progress in Informatics and Computing (PIC), Shanghai, China, 18–20 December 2020; pp. 396–400. [CrossRef]
44. Zhang, C.; Zhen, Z.; Wang, D.; Li, M. UAV path planning method based on ant colony optimization. In Proceedings of the 2010 Chinese Control and Decision Conference, CCDC 2010, Xuzhou, China, 26–28 May 2010; pp. 3790–3792. [CrossRef]
45. Fabien-brulport. Ant Colony Optimisation. Available online: https://github.com/fabien-brulport/ant-colony (accessed on 23 July 2024).
46. Oborne, M. ArduPilot Mission Planner. Available online: https://github.com/ArduPilot/MissionPlanner (accessed on 23 July 2024).
47. Koubaa, A.; Allouch, A.; Alajlan, M.; Javed, Y.; Belghith, A.; Khalgui, M. Micro Air Vehicle Link (MAVlink) in a Nutshell: A Survey. *IEEE Access* **2019**, *7*, 87658–87680. [CrossRef]

48. Zhang, H.; Lan, Y.; Suh, C.P.; Westbrook, J.K.; Lacey, R.; Hoffmann, W.C. Differentiation of Cotton From Other Crops at Different Growth Stages Using Spectral Properties and Discriminant Analysis. *Trans. ASAB* **2012**, *55*, 1623–1630. [CrossRef]
49. Yadav, P.; Thomasson, J.A.; Enciso, J.; Samanta, S.; Shrestha, A. Assessment of different image enhancement and classification techniques in detection of volunteer cotton using UAV remote sensing. In Proceedings of the Autonomous Air and Ground Sensing Systems for Agricultural Optimization and Phenotyping IV, Baltimore, MD, USA, 15–16 April 2019; Volume 11008, pp. 152–165. [CrossRef]
50. Westbrook, J.K.; Eyster, R.S.; Yang, C.; Suh, C.P.C. Airborne multispectral identification of individual cotton plants using consumer-grade cameras. *Remote. Sens. Appl.* **2016**, *4*, 37–43. [CrossRef]
51. Rumora, L.; Miler, M.; Medak, D. Impact of various atmospheric corrections on sentinel-2 land cover classification accuracy using machine learning classifiers. *ISPRS Int. J. Geo-Inform.* **2020**, *9*, 227. [CrossRef]
52. Gan, T.; Zha, Z.; Hu, C.; Jin, Z. Detection of Polyps During Colonoscopy Procedure Using YOLOv5 Network. In Proceedings of the 3rd International Workshop and Challenge on Computer Vision in Endoscopy (EndoCV2021), Nice, France, 13 April 2021.
53. Ma, G.; Haibin, D.; Liu, S. Improved Ant Colony Algorithm for Global Optimal Trajectory Planning of UAV under Complex Environment. *Int. J. Comput. Sci. Appl.* **2007**, *4*, 57–68.
54. Bianchi, L.; Dorigo, M.; Gambardella, L.M.; Gutjahr, W.J. A survey on metaheuristics for stochastic combinatorial optimization. *Nat. Comput.* **2009**, *8*, 239–287. [CrossRef]
55. Ojha, V.K.; Abraham, A.; Snášel, V. ACO for continuous function optimization: A performance analysis. In Proceedings of the 2014 14th International Conference on Intelligent Systems Design and Applications, Okinawa, Japan, 28–30 November 2014; pp. 145–150. [CrossRef]
56. Ebadinezhad, S. DEACO: Adopting dynamic evaporation strategy to enhance ACO algorithm for the traveling salesman problem. *Eng. Appl. Artif. Intell.* **2020**, *92*, 103649. [CrossRef]
57. Barreto, S. Data Augmentation. Baeldung. Available online: https://www.baeldung.com/cs/ml-data-augmentation#:~:text=Data%20Augmentation%20on%20Test,%20Validation,also%20use%20it%20during%20testing (accessed on 9 January 2022).

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

Detecting Hailstorms in China from FY-4A Satellite with an Ensemble Machine Learning Model

Qiong Wu [1,*], Yi-Xuan Shou [2,3,4], Yong-Guang Zheng [5], Fei Wu [1] and Chun-Yuan Wang [1]

[1] School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai 201620, China; wufei@sues.edu.cn (F.W.); wangcy@sues.edu.cn (C.-Y.W.)
[2] National Satellite Meteorological Centre, China Meteorological Administration, Beijing 100081, China; shouyx@cma.gov.cn
[3] Key Laboratory of Radiometric Calibration and Validation for Environmental Satellites, China Meteorological Administration (LRCVES/CMA), Beijing 100081, China
[4] FengYun Meteorological Satellite Innovation Center (FY-MSIC), China Meteorological Administration, Beijing 100081, China
[5] National Meteorological Centre, Beijing 100081, China; zhengyg@cma.gov.cn
* Correspondence: wuqiong8616@sues.edu.cn

Abstract: Hail poses a significant meteorological hazard in China, leading to substantial economic and agricultural damage. To enhance the detection of hail and mitigate these impacts, this study presents an ensemble machine learning model (BPNN+Dtree) that combines a backpropagation neural network (BPNN) and a decision tree (Dtree). Using FY-4A satellite and ERA5 reanalysis data, the model is trained on geostationary satellite infrared data and environmental parameters, offering comprehensive, all-day, and large-area hail monitoring over China. The ReliefF method is employed to select 13 key features from 29 physical quantities, emphasizing cloud-top and thermodynamic properties over dynamic ones as input features for the model to enhance its hail differentiation capability. The BPNN+Dtree ensemble model harnesses the strengths of both algorithms, improving the probability of detection (POD) to 0.69 while maintaining a reasonable false alarm ratio (FAR) on the test set. Moreover, the model's spatial distribution of hail probability more closely matches the observational data, outperforming the individual BPNN and Dtree models. Furthermore, it demonstrates improved regional applicability over overshooting top (OT)-based methods in the China region. The identified high-frequency hail areas correspond to the north-south movement of the monsoon rain belt and are consistent with the northeast-southwest belt distribution observed using microwave-based methods.

Keywords: hail; FY-4A; satellite; machine learning; ensemble learning; feature selection

1. Introduction

Hail, which comprises balls or chunks of ice, descends from vigorously developing cumulonimbus clouds (hailstorms) and reaches the ground with significant momentum, often accompanied by strong winds. Despite its localized impact and short-lived nature, hail has the potential to inflict substantial damage in the context of agriculture, industry, telecommunications, and transportation, thereby posing a significant threat to both lives and property. China, which is heavily influenced by a monsoon climate, is among the regions globally that are frequently affected by hail. The frequency of hail events escalates with the arrival of the summer monsoon in spring, reaching its peak in early summer [1]. In the past, with grain production being low, hail had the potential to result in extensive crop failures, leading to famine and highlighting hail as a particularly devastating weather phenomenon [2]. While advances in meteorological and agricultural technologies have made famines due to hail highly improbable, hail still causes CNY billions in economic losses annually in China [3–12] (see Section 2.1.3 for details). Furthermore, new challenges

have arisen, including hail-induced damage to sophisticated scientific equipment. For instance, the Five-hundred-meter Aperture Spherical radio Telescope (FAST)—a project in China–that was under construction for 22 years—embodies the world's most advanced scientific and technological capabilities. Despite its massive size, it is particularly vulnerable to hail damage. Given that it has over 4000 mirror panels, damage to even a single panel could result in significant losses to scientific research and the economy. Hence, the prevention of hail disasters is of paramount importance. Precise monitoring of hail is an initial and crucial step in disaster prevention and mitigation strategies.

Hailstorms are mesoscale weather systems, characterized by spatial scales of less than several hundred kilometers and durations of a few hours or less. Traditional meteorological observation networks frequently overlook weather systems at this scale, necessitating reliance on remote sensing methods for hail monitoring [13]. Radars provide an unmediated method of hail monitoring through the direct detection of precipitation particles. As such, radar-based hail detection technology was developed early. Witt et al. [14] introduced the Severe Hail Index (SHI) and the Maximum Estimated Size of Hail (MESH), which are calculated through the weighted integration of radar reflectivity data above the melting level, exceeding 40 dBZ. Following the deployment of the WSR-88D radar network, this methodology has been adopted by meteorological stations across China. Wang et al. [15] have introduced an innovative method for early-stage hailstorm identification, utilizing time series analysis on two comprehensive features derived from the transformation of ten radar data geometric properties. However, due to the limited coverage of a single radar, a network of radars is required for extensive monitoring. Variability between radar systems can often result in inconsistent findings. Furthermore, the effectiveness of radar in identifying hail is diminished in mountainous areas such as Guizhou, where the terrain influences radar echoes through blocking or reflection. Additionally, installing and maintaining radar systems in high-altitude plateau regions, such as the Qinghai–Tibet Plateau, poses unique challenges. In conclusion, while radar methods are valuable for hail monitoring, relying solely on them presents challenges, which are particularly evident in the context of effective large-area hail monitoring, as well as in global hail distribution and climatological statistical analysis.

Satellites provide extensive and continuous top-down observations, covering vast areas unimpeded by the underlying surface, enabling the creation of a consistent, long-term, and uniform large-scale hail observation dataset. The technology for detecting hail with the Global Precipitation Measurement (GPM) satellite's microwave imager has reached a relatively mature stage. This technology is based on the principle that large ice hydrometeors—such as graupel or hail—lifted by strong updrafts exhibit extremely low brightness temperatures [16,17]. Ni et al. [18] proposed a 37 GHz polarization-corrected brightness temperature (PCT) threshold below 230 K, which provides the optimal Heidke Skill Score (HSS) for hail identification, as validated against ground reports. Mroz et al. [19] identified a 261 K PCT threshold at 18.7 GHz as having the greatest potential for hail detection within the GPM Microwave Imager's channels, yielding a Critical Success Index (CSI) of 26%. Cecil and Blankenship [20] did not define a specific 37 GHz PCT threshold, but explored how the correlation between 37 GHz PCT and hail reports could be used to estimate the frequency of global hailstorms. Bang and Cecil [21] have developed a hail detection method that integrates probabilities from the minimum 19 GHz PCT and the normalized depression of 37 GHz PCT, utilizing the square root of their product to reconcile inconsistencies when one metric suggests a substantially higher likelihood of hail than the other. Throughout this text, the terms Ni17, Mroz17, CB12, and BC19 refer to the passive microwave retrieval algorithms detailed in the works of Ni et al. [18], Mroz et al. [19], Cecil and Blankenship [20], and Bang and Cecil [21], respectively. While the GPM microwave imager provides advanced hail climate statistics, encompassing global distribution and weather background analysis, its application in hail weather warnings is constrained by the low temporal resolution of its polar orbit, which permits only two daily passes over any specific location.

The new generation of geostationary satellites, boasting enhanced spatial resolution and spectral capabilities, are adept at monitoring rapid and localized severe weather events, such as hail. However, as their infrared channels cannot penetrate clouds, they depend on indirect detection methods using cloud-top brightness temperatures, which are more complex and less developed than direct radar and passive microwave techniques. Punge et al. [22] have employed the detection of overshooting tops (OTs) from geostationary satellite imagery to gauge hail potential, a method pioneered by Bedka et al. [23] and Bedka [24]. Acknowledging the link between OTs and diverse severe weather phenomena, such as heavy rainfall, damaging winds, and tornadoes [25–27], Punge et al. [28] refined this method through incorporating CAPE, 0–6 km bulk wind shear, and freezing level height as filters to eliminate OTs that are unsuitable for hail development in dynamic thermal conditions. Despite these enhancements, the OT-based method is still more effective in equatorial regions and shows considerable geographical variability. Going forward, we refer to the OT-based method and its associated filtering technique as OT and OTfilter, respectively, following the nomenclature established by Punge et al. [22,28]. Merino et al. [29] introduced a daytime hail detection tool that utilizes MSG satellite data, employing a logistic regression algorithm to generate both a convection mask and a hail mask. However, the model's simplicity and linearity lead to a risk of underfitting, especially given the complexities of indirect hail detection. Melcón et al. [30] concentrated exclusively on satellite data, which omits the environmental context which is essential for a comprehensive understanding of hail events. Moreover, despite both studies leveraging visible channels to improve hail detection accuracy, their methods did not achieve an all-day hail detection capability. To maximize the potential of geostationary meteorological satellites in monitoring hail weather, this study aims to develop an advanced, all-day, and large-area hail monitoring model. This model will utilize geostationary satellite infrared channel data, integrated with environmental physical parameters, with a focus on the central and eastern regions of China as the study area.

Hail monitoring, as a research problem, fundamentally presents itself as a binary classification challenge, focused on ascertaining the presence or absence of hail. Machine learning techniques have demonstrated remarkable effectiveness in addressing classification challenges in earth sciences, encompassing methods such as decision trees, random forests, support vector machines (SVMs), K-nearest neighbors (KNN), and neural networks, as illustrated by various studies [31–37]. Sehad and Ameur [38] have introduced a high-accuracy daily rainfall estimation technique that utilizes a multilayer perceptron (MLP) and a multiclass SVM approach. Yang et al. [39] have employed the KNN algorithm to classify convective and stratiform precipitation from ground-based Doppler radar data, showcasing its high precision and effectiveness across diverse precipitation scenarios. Su et al. [40] have applied a suite of machine learning techniques, including logistic regression, random forest, decision trees, and extra trees, to develop a classifier for the rapid intensification of tropical cyclones that outperformed existing forecasting methods. Kim et al. [41] have developed a marine fog detection algorithm that integrates Geostationary Ocean Color Imager (GOCI) observations with Himawari-8 data and employs a decision tree approach, effectively differentiating fog from clouds and offering a high-resolution fog product to enhance marine traffic management. Machine learning offers an innovative and potent tool for enhancing the accuracy of hail predictions. Yao et al. [42] have successfully developed an accurate hail forecasting model for the Shandong Peninsula, leveraging radar data and the random forest algorithm. However, there is still an unmet need for a satellite-based model for hail identification that can be effectively applied across broader regions of China.

This study integrates a backpropagation (BP) neural network with a decision tree model into an ensemble machine learning model for hail detection, thus enhancing the predictive accuracy [40,43]. The premise of ensemble learning is that although individual weak classifiers might struggle with accuracy independently, their combined insights can offset their weaknesses, thereby enhancing the overall accuracy. This approach is especially valuable for hail prediction, given the infrequency of events and the intricate

physical dynamics, which makes it challenging for any single model to fully capture the subtleties [44]. The BP neural network is recognized for its ability to model intricate nonlinear relationships. In this study, a shallow (two-layer) neural network is utilized due to the limited amount of hail data, which can help to mitigate overfitting. Decision tree models are known for their good interpretability; they can automatically select the most informative features for node splitting and are well suited for small datasets. Even with a modest amount of data, they can construct effective models. Through merging these models, the ensemble model leverages the complementary strengths of each, thus boosting the probability of detection (POD) while maintaining a balanced false alarm ratio (FAR). Furthermore, statistical distribution analysis of hail occurrences in China indicates that the decision tree model is more attuned to northern hail events, whereas the BP neural network exhibits greater sensitivity to those in the south. This insight highlights the value of an ensemble learning approach, which is capable of adjusting to regional disparities and delivering a more robust and precise detection framework.

This study aims to introduce a sophisticated, all-day, and large-area hail detection model for China, which employs ensemble learning to integrate a BP neural network and decision tree with geostationary meteorological satellite infrared data. The remainder of this paper is organized as follows: the second section provides an overview of the satellite, reanalysis, and hail data used in this study, as well as the methodology for developing the hail detection model; the third section elaborates on the results of convective cloud identification, feature selection, and model validation, along with the statistical analyses conducted to assess the geographical distribution of hail; the fourth section provides a comprehensive discussion of results, as well as directions for follow-up research; and the final section offers a summary.

2. Data and Methods

2.1. Data

The geographical scope of the study, illustrated in Figure 1, extends from 20° to 50°N in latitude and 105° to 135°E in longitude. Three types of data were leveraged: satellite, reanalysis, and hail. Data from March to August 2018 were used to develop and validate the machine learning models due to the availability of reliable hail event records that are precise to the minute during this period. In contrast, data from March to August 2019 to 2023 were used for the statistical analysis of the spatial distribution of hail and comparative analysis as, during this period, only hail event records precise to the hour were available.

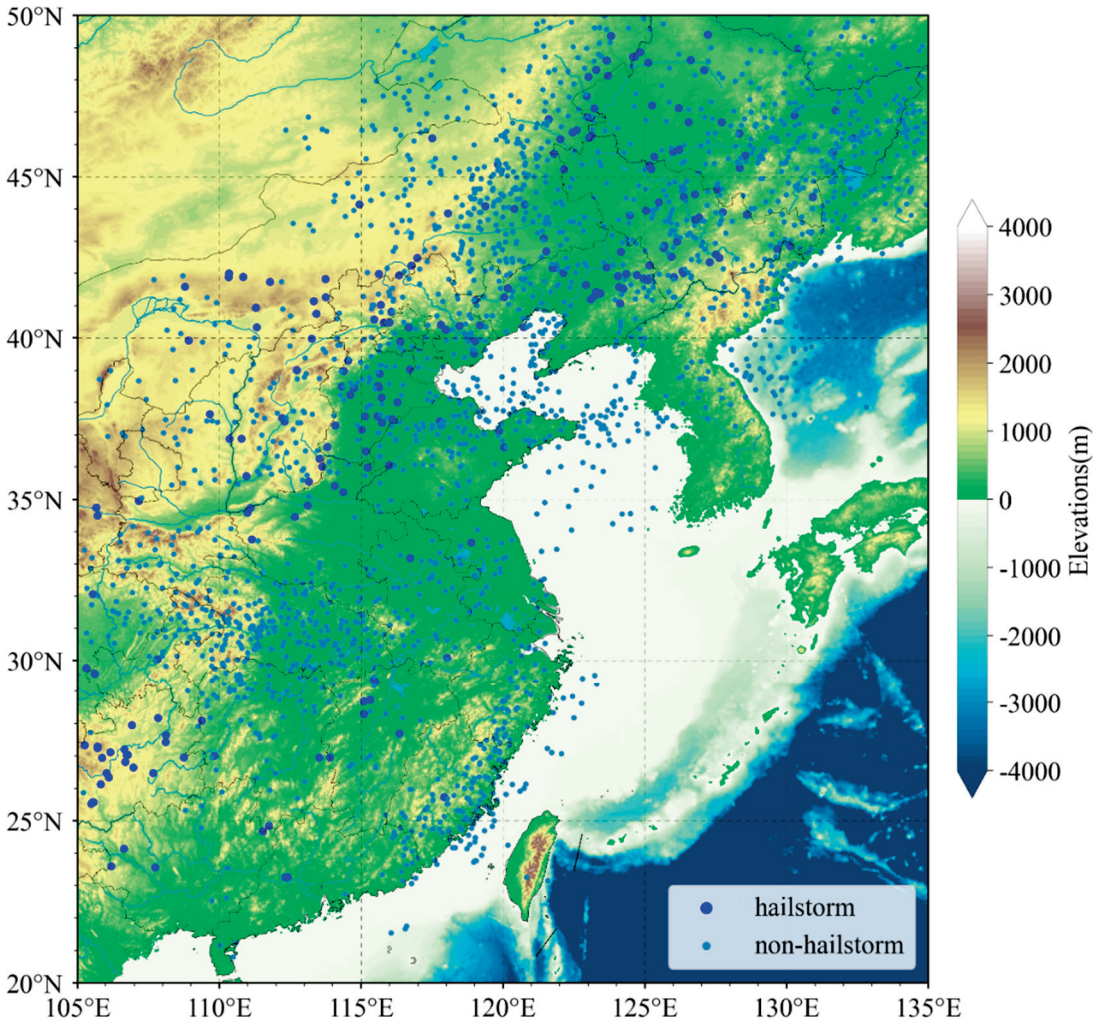


Figure 1. Study area and sample distribution. Dark blue circles indicate hailstorm samples, while light blue circles denote non-hailstorm samples.

2.1.1. Satellite Data

(a) FY-4A

The FY-4A satellite, positioned at 104.7°E longitude above the equator since 25 May 2017, was officially commissioned for use on 25 September 2017. The FY-4A satellite's onboard Advanced Geostationary Radiation Imager (AGRI) is equipped with a total of 14 channels. It provides a spatial resolution ranging from 0.5 to 1 km in the visible/near-infrared band, and from 2 to 4 km in the infrared band. The full disk scan requires 15 min, whereas the scan for the China region takes approximately 5 min. For the development of an all-day hail detection model, this study exclusively uses the brightness temperature data from the six infrared bands at wavelengths of 6.25, 7.1, 8.5, 10.7, 12.0, and 13.5 μm.

(b) GPM PF database

The Global Precipitation Measurement (GPM) [45] Mission Core Observatory satellite was launched on 27 February 2014. The satellite is equipped with instruments including the GPM Microwave Imager (GMI), a passive multi-channel conical-scanning microwave radiometer, and the Dual-frequency Precipitation Radar (DPR), which functions at 13.6 GHz (Ka band) and 35.5 GHz (Ku band). These instruments enable the global measurement of precipitation, ranging from light to heavy. Precipitation Features (PFs), as defined by Nesbitt et al. [46], consist of contiguous pixels meeting specific criteria derived from the swath data of the DPR or GMI. For example, this study uses a PF database known as

'ku40dbzf' [47–49], characterized by areas with Ku band reflectivity exceeding 40 dBZ within the column, which is within the normal swath of Ku. The parameters MIN19PCT and MIN37PCT are employed to define hail PFs, adhering to the methodologies outlined by Ni17, Mroz17, and CB12.

(c) GPM Hail Climatology Data Products

The Passive Microwave Hail Climatology Data Products, as detailed by Cecil and Bang [50], offer gridded ($2^\circ \times 2^\circ$) annual estimates of severe hailstorm frequency. These estimates are based on data from the GPM GMI. The hail probability for each feature is determined using the 37 GHz and 19 GHz PCT retrievals with the BC19 method. A feature is defined as a contiguous area exceeding the size of one GPM GMI pixel, with an 89-GHz PCT value below 200 K. Furthermore, features undergo a screening process to exclude surface artifacts and deep convection, as described by Bang and Cecil [16]. Subsequently, features with an estimated hailstorm probability of at least 20% are tallied within each 2-degree grid box.

2.1.2. Reanalysis Data

The fifth generation of ECMWF reanalysis, ERA5, integrates model data with observations via data assimilation for comprehensive global climate and weather analysis. The data have been re-gridded to a regular latitude–longitude grid with a 0.25-degree resolution. ERA5 provides hourly estimates across a wide range of variables, including atmospheric, oceanic, wave, and land-surface parameters. At varying pressure levels, seven key quantities are used to calculate environmental parameters: geopotential ($m^2\ s^{-2}$), potential vorticity ($K\ m^2\ kg^{-1}\ s^{-1}$), relative humidity (%), specific humidity ($kg^{-1}\ kg^{-1}$), temperature (K), U-component of wind ($m\ s^{-1}$), and V-component of wind ($m\ s^{-1}$). These quantities are crucial for assessing parameters such as the Convective Available Potential Energy (CAPE), the height of 0 °C isotherm, 0–3 km vertical wind shear ($VWS_{0–3}$), Storm-relative helicity (SRH), and the Lapse-rate tropopause (LRT), among others. Additionally, the single-level quantity of skin temperature is used to detect snow- and ice-covered surfaces, as discussed by Zhou et al. [51]. These surfaces can result in significantly low PCT values in the data retrievals, according to Ni17, Mroz17, and CB12 methodologies.

2.1.3. Hail Data

(a) Hail Event Records

Hail events are routinely recorded at weather stations within China's national surface meteorological observational network. Standard hail event records include latitude, longitude, date, exact time of occurrence (down to the hour and minute), and the maximum hailstone diameter. These records are measured or estimated by professionally trained meteorological observers in accordance with observational guidelines, ensuring that they are both detailed and reliable. Data from March to August 2018 were specifically used for training, validation, and testing of the machine learning models.

(b) 1-h Hail Event Records

The other type of hail data used in this study consists of hourly hail event records, which include the latitude, longitude, date, and hour of hail occurrence. These hail event records are automatically captured by instruments and lack precise observation times (i.e., to the minute) and records of the maximum hailstone diameter. 1-h hail event records from March to August from 2019 to 2023 were utilized for the statistical analysis of the spatial distribution of hail occurrences, serving as a benchmark for comparing the various hail detection methods.

(c) Direct Economic Losses

We also retrieved some data from the Yearbook of Meteorological Disasters in China to illustrate the direct economic losses caused by hail and other severe convective weather each year from 2012 to 2021, as shown in Table 1 [3–12].

Table 1. Direct economic losses caused by hail, tornadoes, and other severe convective weather from 2012 to 2021.

Year	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
Direct Economic Losses (in Billion Yuan)	344.1	456.2	276.7	322.7	463.9	147.5	168.3	183.4	282.3	268.7

2.2. Methods

To detect hailstorms from satellite images, the initial step involves identifying convective clouds. Subsequently, an ensemble classifier—which is an 'OR' model combining a BP neural network and a decision tree—is constructed to determine whether the convective cloud is a hailstorm or not. The entire process is shown in Figure 2.

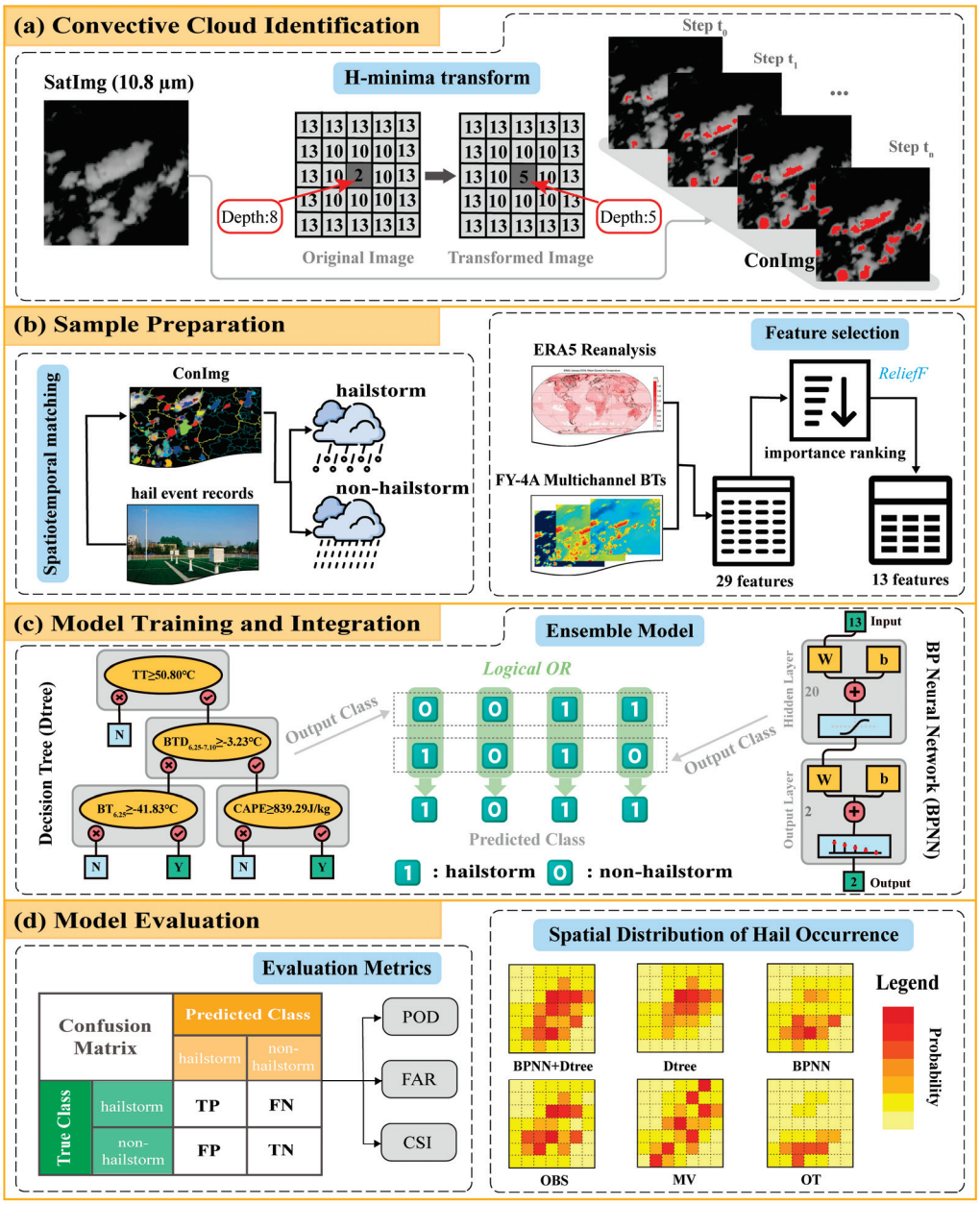


Figure 2. Technical roadmap for developing an ensemble machine learning model for hailstorm identification.

(a) Convective Cloud Identification

On satellite images, convective cloud (CC) areas frequently merge with other cloud regions, making it essential to isolate CCs before analyzing their physical properties. This study employs an iterative H-minima transform algorithm, as referenced in the work of Liu et al. (2014) [52], for this purpose. The algorithm includes five steps per iteration after assigning initial values, as shown in Figure 3a: (1) H-Minima Transform Execution—Conduct an H-minima transform on the satellite image (SatImg). For any connected region S with a minimum brightness temperature $BT_{\min}$, if the BT at any point x within S satisfies $BT(x) \leq BT_{\min} + t$, where t is the H-minima transform threshold (varying from t_0 to t_n with an increase in dt), then adjust $BT(x)$ to $BT_{\min} + t$. Otherwise, retain the original $BT(x)$. This process generates an image, which we denote as HImg. (2) Regional Minima Identification—On HImg, pinpoint regional minima and assign them a value of 1, while setting all other pixels to 0, creating a binary image denoted as BImg. (3) Connected Region Division—Segment the connected regions in BImg into distinct CC areas. (4) CC Image Update—For the convective cloud image (ConImg), if the threshold t is less than t_{merg} (the merging threshold for CCs), set the value of ConImg for each CC (CC_i) to 1. For t greater than or equal to t_{merg}, calculate the intersection count between CC_i and previously identified CCs. If the intersection count is less than 2, set the value of ConImg at CC_i to 1. (5) Iteration Exit Criteria—Decide whether to conclude the iteration process. If t exceeds t_n, terminate the iteration to obtain the final ConImg, where each CC_i denotes a recognized convective cloud. Figure 3b demonstrates several CC identification scenarios: CC_A continues to expand outward with each iteration; CC_B consists of two small CCs that merge into a single CC before $t \geq t_{merg}$; and CC_C and CC_D, due to the two CCs not merging before $t \geq t_{merg}$, remain as two separate CCs in the end. The smaller the value of t_0, the smaller the initial CC, leading to the identification of more CCs; conversely, the larger the value of t_0, the larger the initial CC (merging more small CCs), resulting in fewer CCs identified in the end. The smaller the value of t_n, the smaller the final CC obtained; the larger the value of t_n, the larger the final CC, but it will no longer grow if the peripheral expansion intersects with other CCs. The value of dt determines the speed at which the CC expands outward, while t_{merg} determines whether it is permissible to merge two or more CCs identified in the previous iteration step into one cloud mass identified at this iteration. Merging is only allowed when $t < t_{merg}$. In this study, the values of t_0, t_n, dt, and t_{merg} were set as 1.0, 24, 0.1, and 2.0, respectively.

(b) Sample Preparation

To correlate identified convective clouds (CCs) with hail observation records from meteorological stations between March and August 2018, we employed a spatial and temporal matching process. The matching criteria were as follows: the time difference with the hail observation must not exceed 5 min, there should be intersecting pixels within a 5×5 pixel window surrounding the station, and the CC with the shortest average distance to the station is designated as a hailstorm. Given that hail events are rare, in order to minimize the inclusion of non-hailstorm samples and concentrate on those with high similarity, non-hailstorm samples were selected from CCs at the same moment. However, these samples had to adhere to the criterion of having no hail observation records within a 100-pixel range around them within an hour. Through this matching process, we obtained a dataset comprising 221 hailstorm and 2271 non-hailstorm samples, as illustrated in Figure 1.

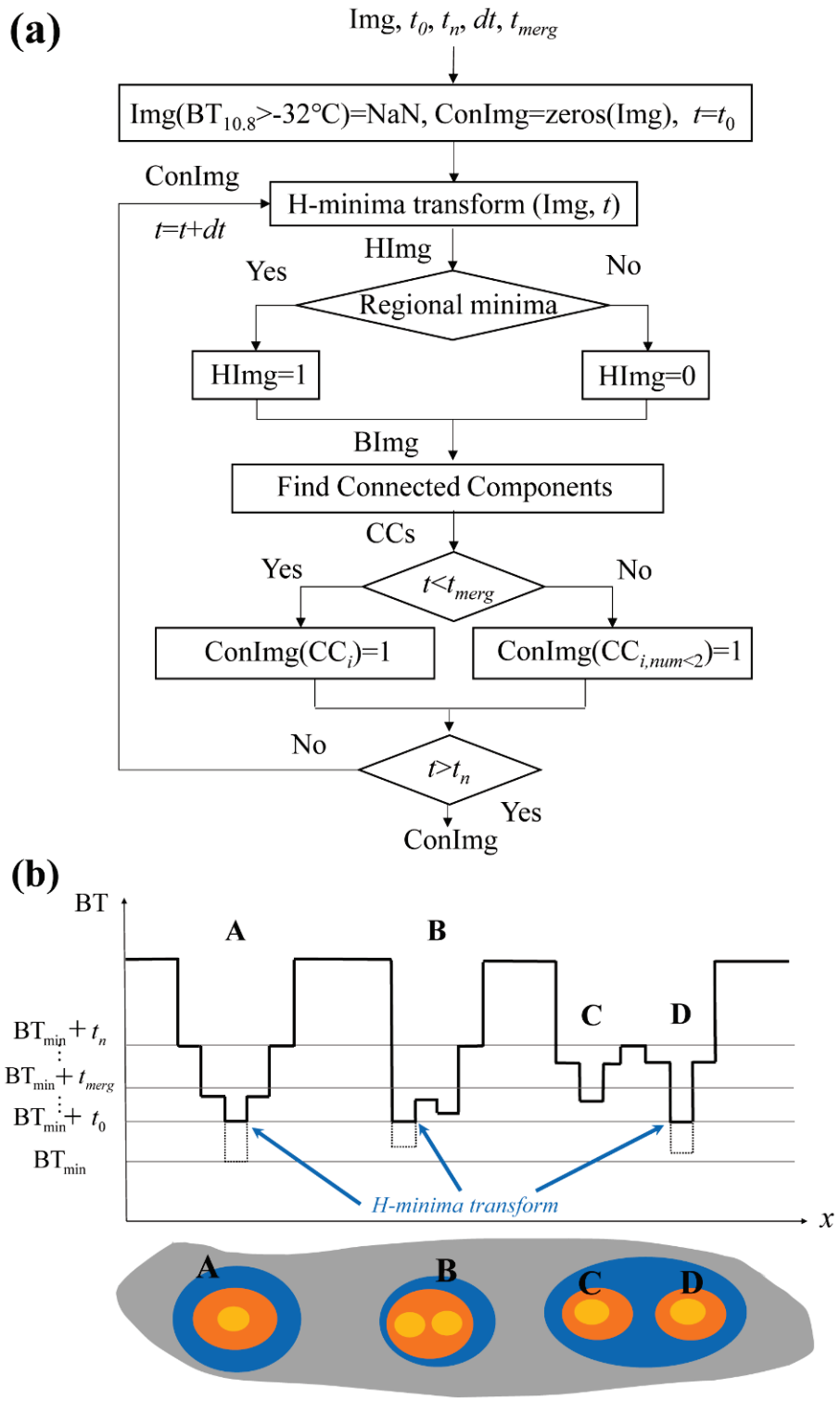


Figure 3. (**a**) Flowchart of the iterative H-minima transform method for identifying convective clouds. (**b**) Schematic of the process for identifying convective clouds with the iterative H-minima transform method.

For each hailstorm and non-hailstorm sample, we calculated 29 physical quantities that are commonly referenced in the literature, as listed in Table 2. These quantities include cloud-top brightness temperature (BT), differences between brightness temperatures (BTD), and environmental thermodynamic and dynamic parameters. Detailed information about these quantities, their references, and their physical significance can be found in the work

of Wu et al. [53]. To calculate the 29 physical quantities of hailstorms (non-hailstorms), the values at each pixel within the hailstorm (non-hailstorm) boundary were first estimated. The cloud-top BT/BTD for each pixel can be directly calculated from the satellite infrared channel observations, while the environmental thermodynamic and dynamic parameters for each pixel are derived from the corresponding ERA5 grid data at the pixel's location. The final physical quantities are the average values across all pixels within a given bounded region.

Table 2. Physical quantities calculated for each hailstorm (non-hailstorm) sample.

Parameters	Abbreviations	Units
BT of 6.25-μm channel	$BT_{6.25}$	°C
BT of 7.10-μm channel	$BT_{7.10}$	°C
BT of 8.50-μm channel	$BT_{8.50}$	°C
BT of 10.8-μm channel	$BT_{10.8}$	°C
BT of 12.0-μm channel	$BT_{12.0}$	°C
BT of 13.5-μm channel	$BT_{13.5}$	°C
BTD between 6.25- and 10.8-μm channels	$BTD_{6.25–10.8}$	°C
BTD between 7.10- and 10.8-μm channels	$BTD_{7.10–10.8}$	°C
BTD between 6.25- and 7.10-μm channels	$BTD_{6.25–7.10}$	°C
BTD between 8.50- and 10.8-μm channels	$BTD_{8.50–10.8}$	°C
BTD between 12.0- and 10.8-μm channels	$BTD_{12.0–10.8}$	°C
BTD between 13.5- and 10.8-μm channels	$BTD_{13.5–10.8}$	°C
Convective available potential energy	CAPE	$J\ kg^{-1}$
Convective inhibition	CIN	$J\ kg^{-1}$
Precipitable water	PW	cm
K index	KI	°C
Total totals	TT	°C
Showalter index	SI	°C
Lifted index	LI	°C
Boyden index	BI	unitless
Height of 0 °C	$H_{0\ °C}$	km
Height of −20 °C	$H_{-20\ °C}$	km
Wet-bulb zero height	WBZ	km
Hail growth zone (−10 to −30 °C) thickness	HGZ	km
0–6 km vertical wind shear	$VWS_{0–6}$	$m\ s^{-1}$
0–3 km vertical wind shear	$VWS_{0–3}$	$m\ s^{-1}$
Storm-relative helicity	SRH	$m^2\ s^{-2}$
Bulk Richardson Number	BRN	unitless
Mean high-level (400 to 200 hpa) potential vorticity	PVH	$K\ m^2\ kg^{-1}\ s^{-1}$

We utilized the ReliefF algorithm—as introduced by Kononenko et al. [54] and further developed by Robnik-Šikonja and Kononenko [55]—to rank the significance of these 29 physical quantities in distinguishing between hailstorm and non-hailstorm samples. The ReliefF algorithm allocates weights to features based on their capacity to differentiate samples that are in close proximity to one another within their respective classes. A higher weight indicates that the feature is more effective in distinguishing between the two categories. The ranking outcomes are presented in Figure 4, which reveals six distinct breakpoints in the weight distribution chart. For this study, we chose the breakpoint at the fourth division as the cutoff point. Consequently, we selected the top 13 features, as indicated by their ranking, to serve as input features for our machine learning model.

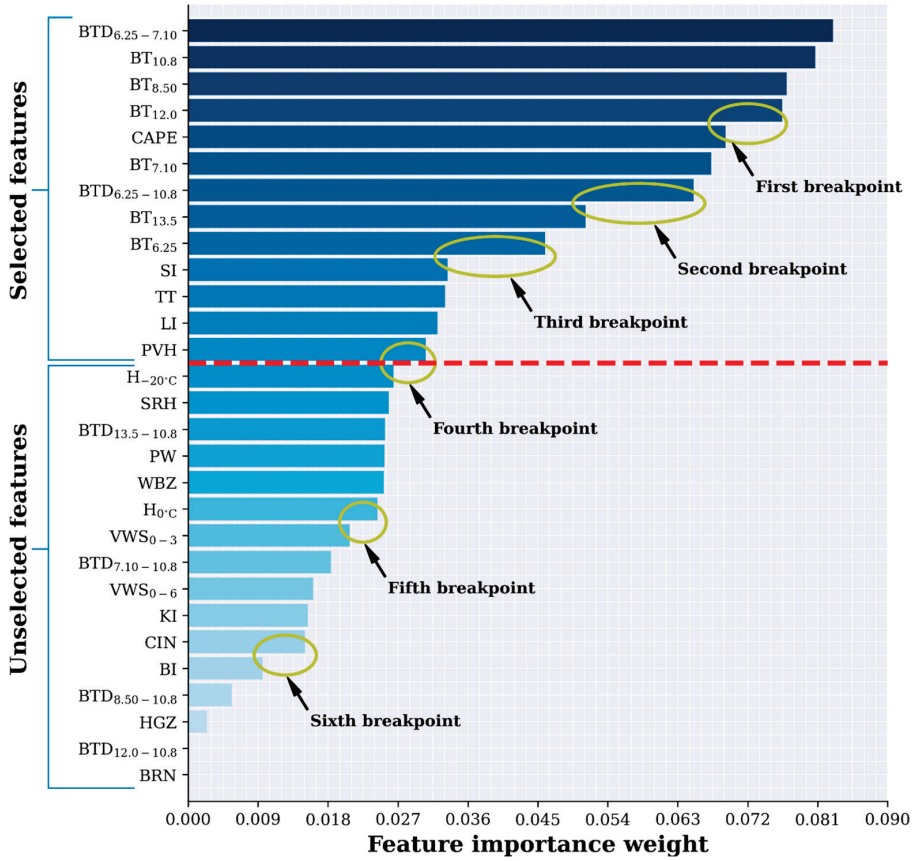


Figure 4. Bar chart of feature importance ranking using the ReliefF method. Six distinct breakpoints are indicated by yellow ellipses. The red dashed line drawn at the fourth breakpoint serves as the cutoff point for this study; features with importance scores above this line (amounting to 13 features) were chosen as inputs for the model, whereas those with lower scores were excluded.

(c) Model Training and Integration

The hailstorm and non-hailstorm samples were divided into three distinct groups: a training set, a validation set, and a test set, with respective proportions of 60%, 10%, and 30%. The training set served as the foundation for computing gradients and updating the weights and biases of the BP neural network (BPNN) model, following the principles outlined by Glorot and Bengio [56]. For the decision tree (Dtree) model, this dataset was instrumental in splitting tree nodes, continuing until the tree reaches its maximum depth or the number of samples in a node falls below a predefined threshold, as described by Loh [57]. The validation set plays a crucial role in identifying the optimal hyperparameters for the models and in curbing overfitting. In the case of the BPNN model, early stopping was implemented as a technique to prevent overfitting. For the Dtree model, the practice of pruning was applied with the same intent. During the model training for hyperparameter selection, grid search was primarily utilized. The selected hyperparameters included the Maximal Number of decision Splits (MaxNumSplits) and Minimum size of leaf node observations (MinLeafSize) for the Dtree model and Hidden Layer sizes (HLayersize) for the BPNN model. Furthermore, given the imbalance in the sample sizes between hailstorm and non-hailstorm samples, and recognizing the serious consequences of misclassifying a hailstorm as a non-hailstorm, we introduced the hyperparameter of Prior Probability (PP) for each class during Dtree model training. This adjustment in class weights reduces the model's bias towards the majority class of non-hailstorms and enhances the detection rate of the minority class of hailstorms. Additionally, when training the BPNN model, a hyperparameter of Error Weights (EW) for each class was incorporated. This assigns a higher weight to the error of misidentifying a hailstorm as a non-hailstorm and a lower

weight to the reverse misidentification, thus aligning the model's outputs with the severity of different misclassification scenarios. The search range and values of the hyperparameters are presented in Table 3. The test set was reserved for the evaluation of model performance, as detailed in step (d).

Table 3. The hyperparameters of the Dtree and BPNN models.

Model	Hyperparameter	Search Range	Value
Dtree	MaxNumSplits	6	6
	MinLeafSize	1:1:30	19
	PP	1 for hailstorms	1
		0.1:0.05:0.5 for non-hailstorms	0.25
BPNN	HLayersize	20:10:40	20
	EW	1 for hailstorms	1
		0.1:0.1:0.5 for non-hailstorms	0.4

The architecture of the trained Dtree and BPNN models are depicted in Figure 2c. The Dtree model was structured with a single root node (TT $\geq$ 50.80 °C), one internal node in the first layer ($BTD_{6.25-7.10} \geq -3.23$ °C), two internal nodes in the second layer ($BT_{6.25} \geq -41.83$ °C and CAPE $\geq$ 839.29 J/kg), and five leaf nodes, which use Yes (Y) or No (N) to indicate whether the final decision is a hailstorm. The BPNN model was designed with a two-layer fully connected architecture, which includes one hidden layer comprising twenty neurons and one output layer with two neurons. The hyperbolic tangent function was used as the activation function of the hidden layer, and the model adopted the cross-entropy loss function.

To bolster the identification of hail hazards, which can be challenging for machine learning models due to their minority class status, an ensemble approach using the logical 'OR' operation to combine the BPNN and Dtree models is proposed, denoted as BPNN+Dtree. Specifically, for any given CC sample, if either the Dtree or the BPNN model predicts it as a hailstorm (assigned a value of 1), the sample is classified as a hailstorm. On the other hand, the sample is classified as a non-hailstorm (assigned a value of 0) only if both models concur on this classification, as depicted in Figure 2c. As the integration is performed on the classification results (not the output probabilities) of the two models, there are only two methods: 'OR' and 'And'. When the 'And' method is used, a sample is considered a hailstorm only if both models identify it as such. However, hailstorms, on one hand, are not easily detected by machine learning models due to their nature as rare occurrences. On the other hand, hail is hazardous weather, and the risk of failing to predict a hailstorm is much greater than that of issuing a false alarm. Therefore, the aim is to identify as many hailstorms as possible. Therefore, based on these two considerations, this study employed the 'OR' method for integration.

(d) Model Evaluation

The model evaluation focused on two main aspects, as shown in Figure 2d. The first aspect involved calculating key evaluation metrics, including POD, FAR, and CSI, all of which were derived from the confusion matrix of the test set. The respective calculation formulas are as follows:

$$\mathrm{POD} = \frac{\mathrm{TP}}{\mathrm{TP+FN}},$$

$$\mathrm{FAR} = \frac{\mathrm{FP}}{\mathrm{TN+FP}},$$

$$\mathrm{CSI} = \frac{\mathrm{TP}}{\mathrm{TP+FP+FN}}.$$

Here, TP (True Positive) denotes the number of samples correctly identified as the hail class, FP (False Positive) is the number of samples incorrectly classified as the hail class when they do not belong to it, FN (False Negative) is the number of hail samples that were incorrectly classified as non-hail, and TN (True Negative) is the number of non-hail samples correctly identified.

The evaluation also included a statistical comparison of the spatial distribution characteristics of hail occurrences, organized into three comparative groups: (1) A comparative analysis of different machine learning models, including the BPNN, Dtree, and BPNN+Dtree models; (2) the ensemble model BPNN+Dtree was compared with the OT and OTfilter hail identification methods; and (3) the ensemble model BPNN+Dtree was also compared with various microwave-based hail identification methods—namely, Ni17, CB12, Mroz17, and BC19. Each hail identification method identifies different objects: machine learning models classify CCs, OT-based methods analyze cloud pixels, and microwave-based methods detect PFs. Additionally, the temporal scope of the statistical analysis varies. The statistical period for machine learning models and OT-based methods spans from 2019 to 2023, while microwave-based methods cover a longer period from 2014 to 2023. This extended period for microwave-based methods is necessary due to the polar orbit of the GPM satellite, which restricts the visibility of spatial distribution characteristics when observed over shorter durations. To standardize the comparison, the hail occurrences were tallied within each $2^\circ \times 2^\circ$ grid cell. Specifically, for the machine learning models, the number of hailstorm center points falling within each grid cell was counted; for the OT-based method, the number of hail cloud pixels falling within each grid cell was calculated; and, for microwave-based methods, the number of hail PFs center points falling within each grid cell was accumulated. These counts were then normalized through dividing them by the total number of hail occurrences, thereby converting the raw counts into probabilities for further analysis.

3. Results

3.1. Results of Convective Cloud Identification

Figure 5a–h depict the results of the iterative H-minima transform method for identification of CCs on the FY-4A 10.8-μm infrared image at 06:38 UTC on 11 August 2018. As the process unfolds, it can be observed that the area of the identified CCs progressively expands with each iteration, as shown in Figure 5a–h. The final identification result is presented in Figure 5h, where different colors represent distinct CCs. At this stage, the area of each CC either ceases to grow or, if it continues to expand, it merges with other CCs. This highlights the dynamic nature of the identification process and the method's effectiveness in delineating CCs.

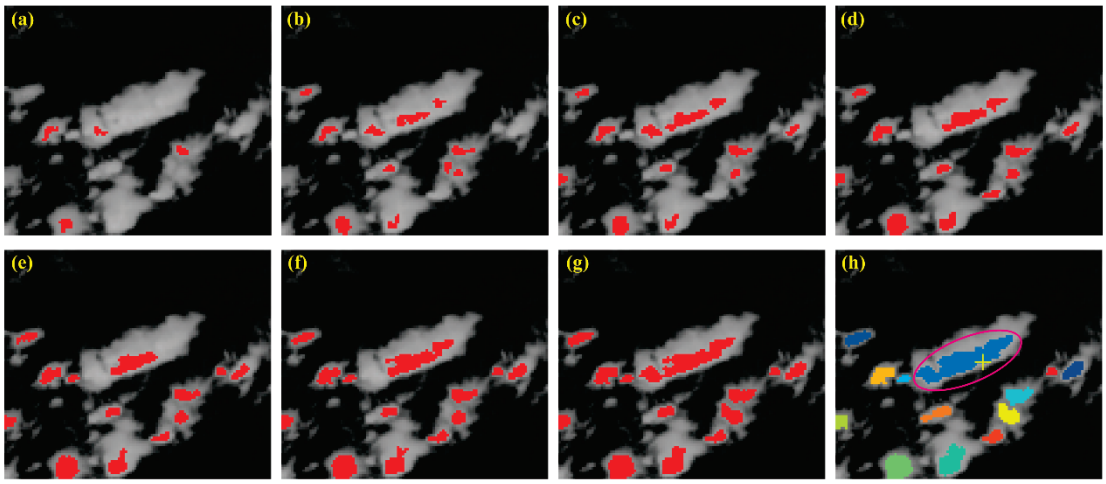


Figure 5. The results of the identification of convective clouds using the iterative H-minima transform method on the FY-4A 10.8-μm infrared channel image, recorded at 06:38 UTC on 11 August 2018. (**a**–**h**) illustrate the outputs at various time steps throughout the process. In (**h**), the identified convective clouds are differentiated by color, with yellow plus signs indicating the positions of observed hail, and magenta ellipses outlining the corresponding hailstorms.

3.2. Statistical Results of Selected and Unselected Features

To validate the feature selection, joy plots for both selected and unselected features were created, as shown in Figures 6 and 7. The blue and orange curves in these figures represent the Gaussian kernel density estimation for the hailstorm and non-hailstorm classes, respectively. Figure 6 demonstrates a clear distinction between the joy plots of hailstorms and non-hailstorms, characterized by distant peaks and distinctly different curve shapes. In contrast, Figure 7 reveals a significant overlap in the joy plots for both classes, with centers and curve shapes showing various degrees of overlap.

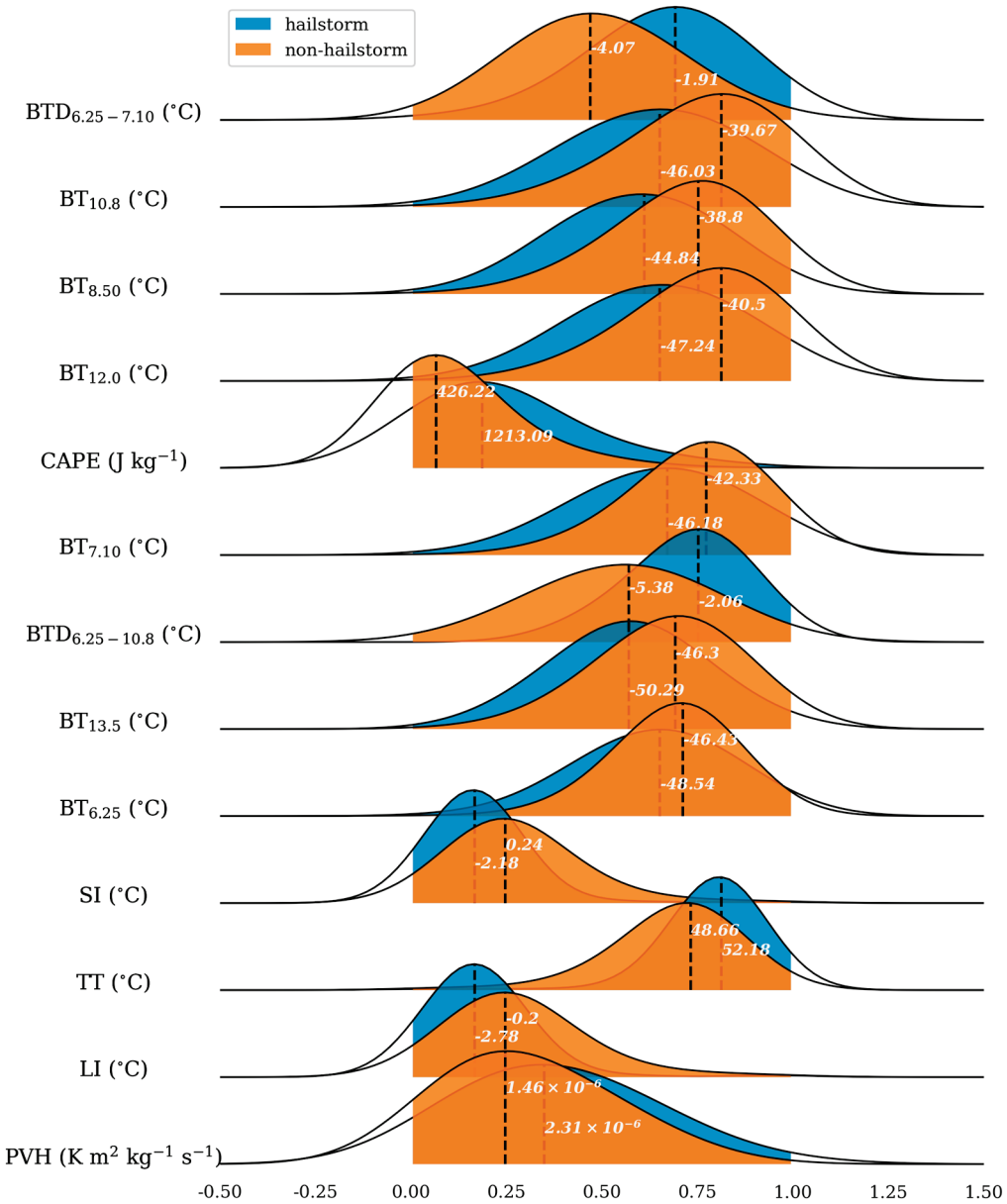


Figure 6. Joy plot of selected features. The curve filled with blue color represents the Gaussian kernel density estimation for hailstorm samples, while the orange curve represents the same for non-hailstorm samples. The horizontal axis represents the normalized value of each feature's physical quantity. The black dashed line marks the feature value where the probability density is at its maximum. The white numbers correspond to the actual physical values.

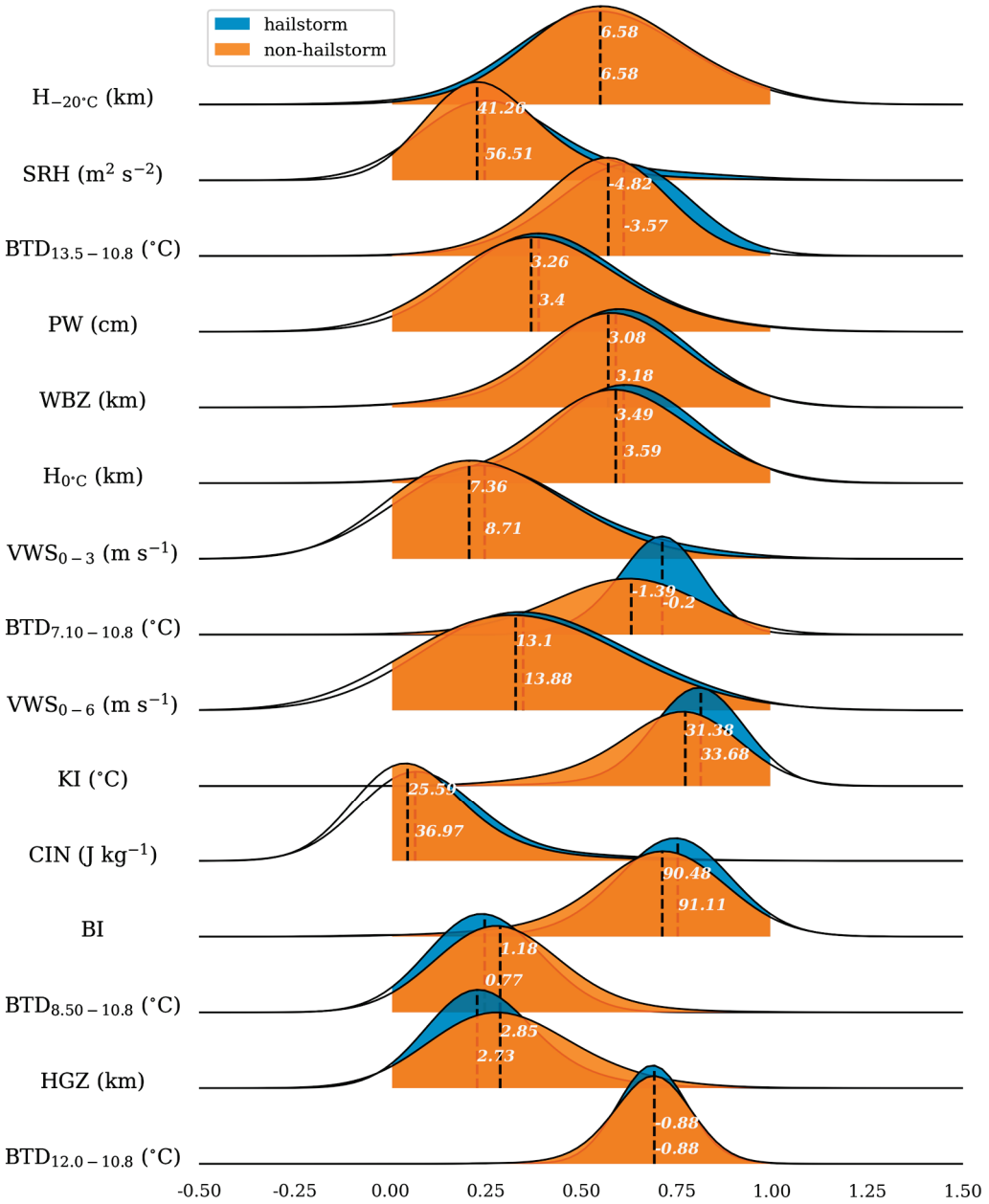


Figure 7. Joy plot for unselected features. Similar to Figure 6, this plot presents the Gaussian kernel density estimation curves for hailstorm and non-hailstorm samples, but for features that were not selected.

The results suggest that the selected features offer a more discernible differentiation between hailstorms and non-hailstorms than the unselected features, making them more rational for use. The selected features mainly encompass physical quantities that reflect cloud-top characteristics, such as height, thickness, and water vapor distribution (e.g., $BTD_{6.25–7.10}$, $BT_{6.25}$), as well as thermodynamic conditions (e.g., CAPE, TT). Dynamic factors such as $VWS_{0–3}$ and SRH were not selected. This selection deviates from previous criteria for hail intensity classification, which prioritized physical quantities related to cloud-top height, dynamic properties, and melting level height, including vertical wind shear, a factor known for its strong correlation with large hail events [53].

3.3. Confusion Matrix Evaluation Results

To assess the model's predictive performance, confusion matrices for the BPNN, Dtree, and BPNN+Dtree models were constructed for the test set, as shown in Figure 8. The evaluation metrics POD, FAR, and CSI were computed and presented in Table 4. From Figure 8, it can be seen that through addressing the sample imbalance issue, the detection rate of hailstorms (minority class) with the three models improved, rather than obtaining a high overall accuracy rate ($\frac{TP+TN}{TP+TN+FP+FN}$) by tendentiously identifying all samples as non-hailstorms (majority class). Table 4 illustrates that the BPNN model consistently outperformed the Dtree model across all evaluated metrics: POD, FAR, and CSI. Notably, the ensemble model BPNN+Dtree had significantly improved POD (reaching 0.69), compared to the individual models. Meanwhile, its FAR value was between those of the BPNN and Dtree models. This indicates that the BPNN+Dtree model achieved a balance that effectively increases the POD for hailstorms without causing a substantial increase in the FAR.

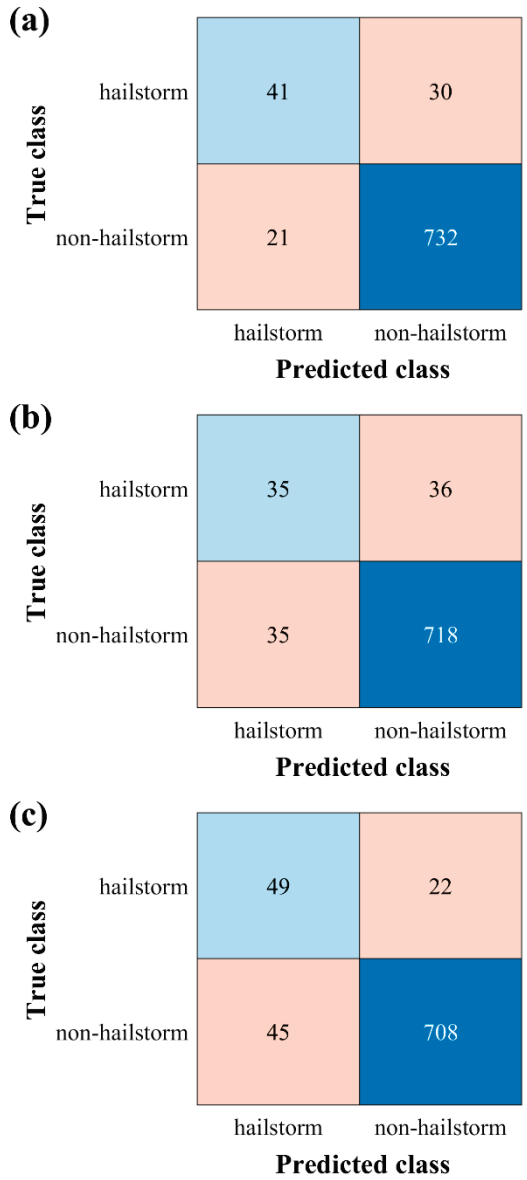


Figure 8. Confusion matrices for various models on the test set. (**a**) BPNN; (**b**) Dtree; (**c**) BPNN+Dtree.

Table 4. Performance metrics (POD, FAR, and CSI) of different classification machine learning models on the test set.

	POD	FAR	CSI
BPNN	0.58	0.34	0.45
Dtree	0.49	0.50	0.33
BPNN+Dtree	0.69	0.48	0.42

3.4. Statistical Results of Hail Occurrence Probability Spatial Distribution

3.4.1. Dtree vs. BPNN vs. BPNN+Dtree vs. OBS

Figure 9a–c presents the spatial distribution of hail occurrence probability calculated based on hailstorms identified from 2019 to 2023 with the three distinct machine learning models. Specifically, Figure 9a illustrates the results for the BPNN model, Figure 9b shows those for the Dtree model, and Figure 9c shows those for the BPNN+Dtree ensemble model. For reference, Figure 9d depicts the spatial distribution derived from 1-h hail event records within the same period.

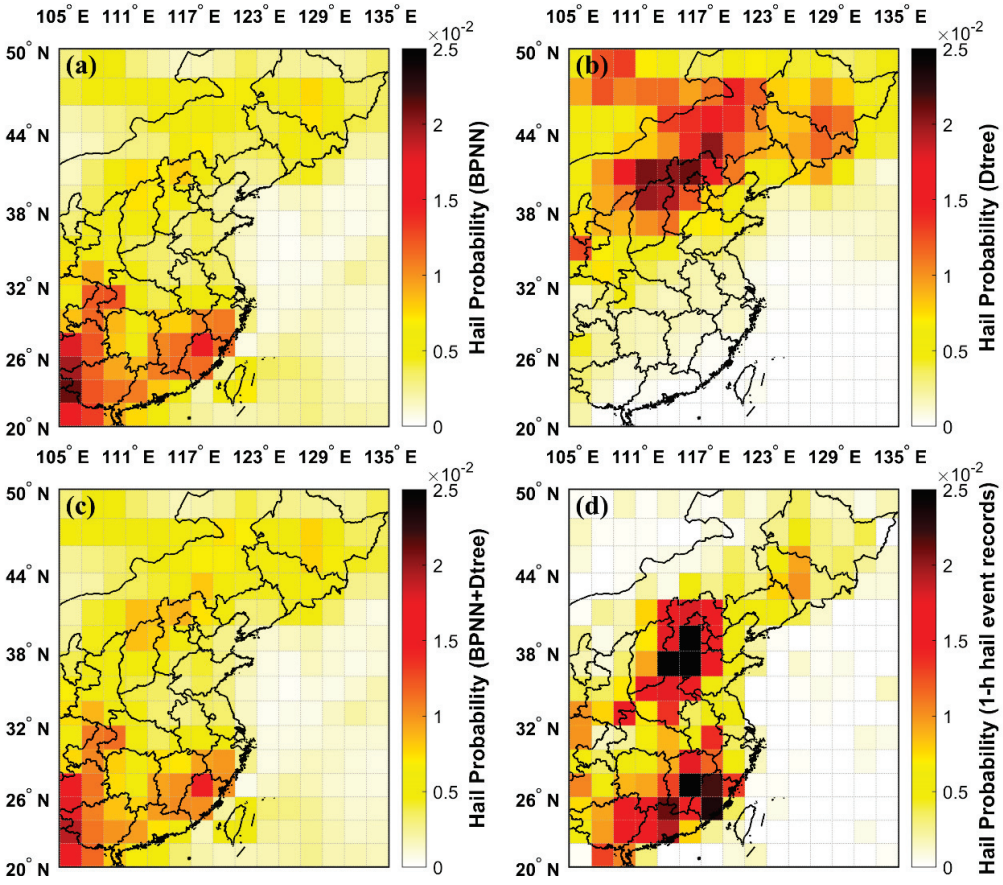


Figure 9. Spatial distribution of hail occurrence probability calculated using various machine learning models: (**a**) The BPNN model; (**b**) the Dtree model; (**c**) the BPNN+Dtree ensemble model; and (**d**) 1-h hail event records.

The BPNN model (Figure 9a) presented a closer match to the observational data (Figure 9d) in terms of spatial distribution characteristics than the Dtree model (Figure 9b). Both the BPNN and observational calculations identified three high-frequency hail regions: the southeast coastal area, North China, and Northeast China. However, the BPNN model was more effective in identifying hail in the southern regions while performing relatively poorly for northern hail. In contrast, the Dtree model was more sensitive to hail in the northern regions. The BPNN+Dtree ensemble model (Figure 9c) capitalizes on the strengths

of both, enhancing the detection capability for hail across the entire region and providing a more accurate spatial distribution of hail probability that closely mirrors the observations. However, as the number of hailstorms identified by the Dtree model was much smaller than that identified by the BPNN model, the BPNN+Dtree ensemble model still had a relatively low probability in the north. Additionally, the machine learning models analyzed hailstorms using satellite infrared images, while the observational data were reliant upon on-site observations. This discrepancy explains the broader spatial distribution range evident in Figure 9a–c, compared to the more focused distribution in Figure 9d.

3.4.2. BPNN+Dtree vs. OT vs. OTfilter vs. OBS

The spatial distribution of hail occurrence probability calculated based on hail cloud pixels identified by the OT method was predominantly concentrated in the tropical region, as shown in Figure 10a. However, when environmental filtering factors were taken into account, the identification results obtained with the OTfilter method were skewed towards the overly northern region. This suggests that the applicability of both the OT and OTfilter methods is relatively limited in the China region. In contrast, the result of the BPNN+Dtree ensemble model presented in Figure 10c, showed a significantly higher degree of congruence with the 1-h hail event records depicted in Figure 10d. The above comparative results indicate that the BPNN+Dtree model has stronger applicability over China.

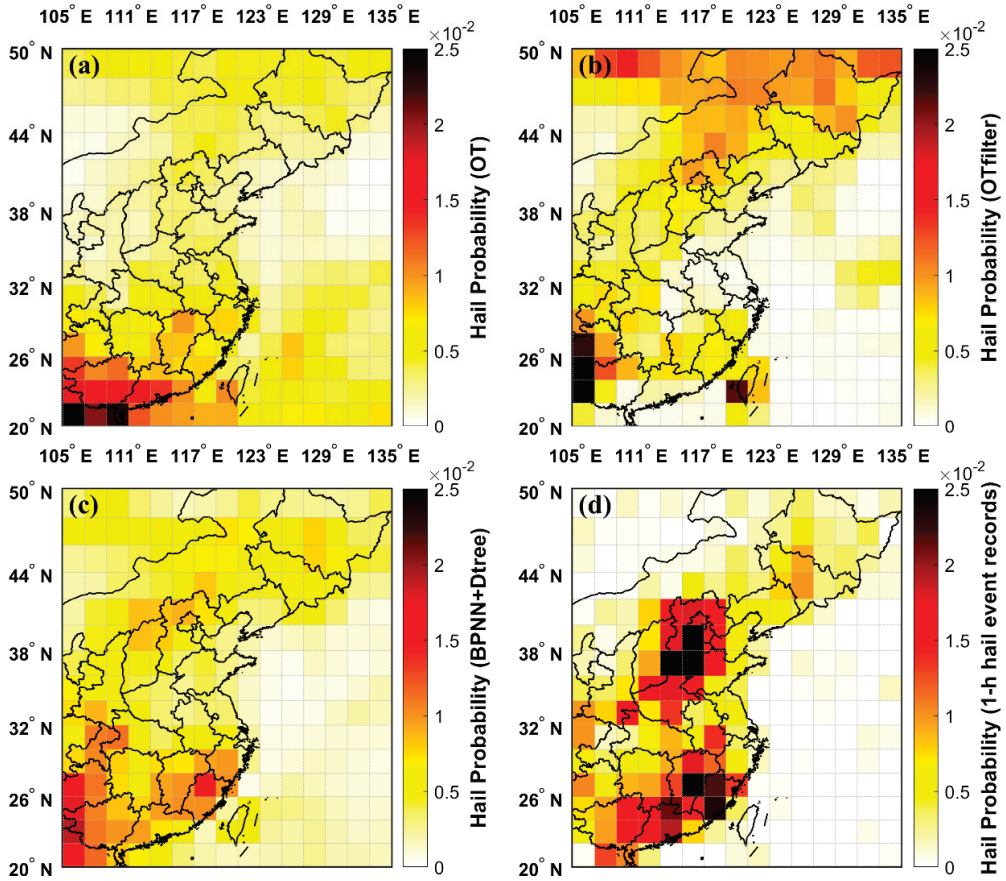


Figure 10. Comparison of the spatial distribution of hail occurrence probability obtained with the BPNN+Dtree ensemble model and OT-based hail identification methods: (**a**) Calculated based on hail cloud pixels identified by the OT method; (**b**) calculated based on hail cloud pixels identified by the OTfilter method; (**c**) calculated using the BPNN+Dtree ensemble model-identified hailstorms; (**d**) calculated based on 1-h hail event records.

3.4.3. BPNN+Dtree vs. Ni17 vs. Mroz17 vs. CB12 vs. BC19 vs. OBS

Figure 11b–e depict the spatial distributions of hail occurrence probability as determined using the microwave-based methods Ni17, Mroz17, and CB12, as well as the reference method BC19. No high-frequency centers were evident, but a clear belt of hail occurrences stretches from the northeast to the southwest, aligning with the boundary of the subtropical high. The distribution pattern of the BPNN+Dtree ensemble model, as shown in Figure 11a, closely aligns with those of the Ni17 and BC19 methods. The model's three high-frequency areas are located on this belt, sequentially forming from south to north. This formation is due to the northward shift of the monsoon rain belt, which follows the northward movement of the subtropical high. It can be seen that the hail distribution characteristics delineated by the BPNN+Dtree ensemble model are in accordance with the actual climatic patterns.

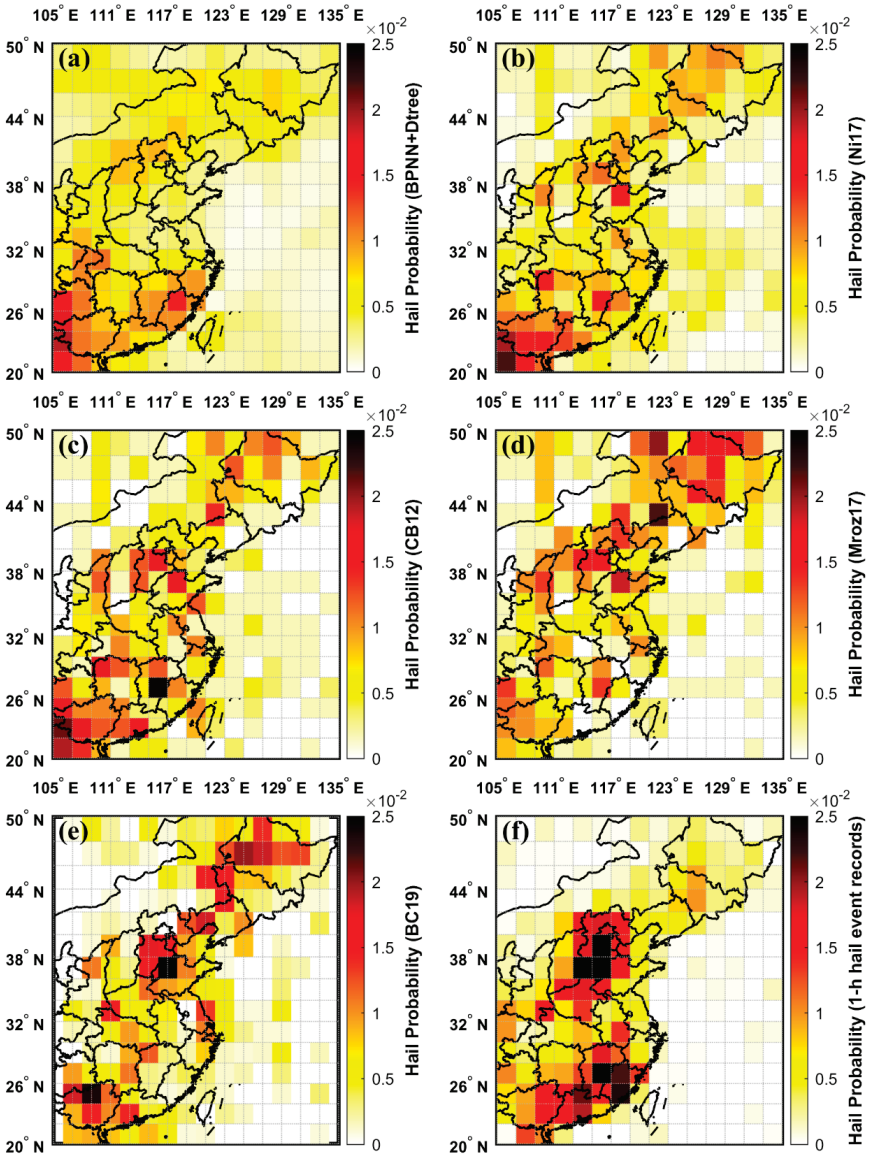


Figure 11. Comparison of the spatial distribution of hail occurrence probability obtained with the BPNN+Dtree ensemble model and microwave-based hail identification methods: (**a**) Calculated using the BPNN+Dtree ensemble model-identified hailstorms; (**b**) calculated using the Ni17 method-identified hail PFs; (**c**) Calculated using the CB12 method-identified hail PFs; (**d**) calculated using the Mroz17 method-identified hail PFs; (**e**) calculated based on the published BC19 annual average data; and (**f**) calculated based on 1-h hail event records.

4. Discussion

This study utilizes FY4A satellite data and ensemble learning models to identify hailstorms in China, thus improving the accuracy of real-time monitoring and early warning systems for hail disasters.

The study began by employing the H-minima transform method to identify CCs, which are then correlated with the hail event records from March to August 2018. This process resulted in the collection of 221 hailstorm samples and 2271 non-hailstorm samples. For each sample, 29 features were calculated, encompassing cloud-top BT (BTD), thermodynamic, and dynamic features. The ReliefF method was then applied to filter these features, leading to the selection of 13 input features that can effectively distinguish between hailstorms and non-hailstorms. A comparative joy plot analysis of the selected versus unselected features revealed that the Gaussian kernel density curves of the selected features showed a greater degree of separation between the two sample types, indicating their superior discriminative power. From a physical perspective, the preferred quantities for classifying hailstorms are those that reflect cloud-top characteristics such as height, thickness, and water vapor distribution (e.g., $BTD_{6.25–7.10}$ and $BT_{6.25}$), as well as thermodynamic conditions such as CAPE and TT. In contrast, for the estimation of hail intensity, the focus is on cloud-top height and dynamic features, such as wind shear, which is a strong indicator of large hail events [53].

Next, Dtree and BPNN models were trained for the classification of hailstorms and non-hailstorms, using prior probability and error weight hyperparameters to improve the detection rate of the hailstorm class (minor class) and using pruning and early stopping techniques to mitigate overfitting. Given the higher risk associated with failing to report hail events compared to the risk of false alarms, the outputs of the Dtree and BPNN models were integrated using a logical 'OR' operation; therefore, the ensemble model—termed BPNN+Dtree—predicts a hailstorm if either model outputs a positive result. The BPNN+Dtree model achieved a POD value of 0.69 on the test set, which is significantly higher than those of the BPNN model (0.58) and the Dtree model (0.49), with an FAR balanced between those of the two models.

To evaluate the long-term identification effectiveness of the BPNN+Dtree model, a statistical analysis of hailstorms identified by the BPNN+Dtree ensemble model from 2019 to 2023 was conducted. The spatial distribution of hail occurrence probability was compared with the results obtained with other methods, including individual Dtree and BPNN models, OT-based identification methods, and microwave-based identification methods. The analysis revealed that the Dtree model is more likely to identify northern hailstorms, while the BPNN model is more likely to identify southern hailstorms. As such, the integration of these two models into the BPNN+Dtree model improved its identification capability across the whole of China. The three high-incidence areas of hail identified—namely, the southeast coastal region, North China, and Northeast China—were in close agreement with observational data. In contrast, the OT-based hail identification method calculated hail high-frequency areas that significantly deviated from the observed patterns tending to concentrate in tropical regions or the northern areas, making it less suitable for hail detection in China. The spatial distribution of hail occurrence probability calculated using the microwave-based methods presented no obvious high-frequency centers. Instead, a northeast–southwest belt distribution was observed, which aligns with the direction of the subtropical high's edge. The hail high-frequency areas calculated with the BPNN+Dtree ensemble model were also located on this belt, forming in sequence from south to north corresponding to the northward shift of the monsoon rain belt following the northward movement of the subtropical high.

Although the above results indicate that this study has established a relatively effective hailstorm detection model based on satellite data using ensemble machine learning methods, there are still aspects that can be further improved. First, the number of hailstorm samples should be increased. Typically, for machine learning models, the more and richer the training samples, the better the generalization performance of the model. Due to the

lack of precise, minute-level reliable hail event records after 2018, the current work only used data from 2018 for training, which is insufficient for establishing a model with strong generalization capabilities. In the future, we plan to combine disaster records, radar observations, and other multi-source data to manually create more reliable hailstorm samples. This will be used for training more sophisticated models, thus improving the accuracy of hail identification. Second, although the ensemble method improved upon the predictive performance of the individual models, to a certain extent, the final results still presented a certain gap with respect to the observations, especially in the identification of hail in the northern regions, which requires further improvement. Based on increasing the training data, leading to improvement of the detection performance of individual models, the ensemble model can be expected to also perform better. Third, the logical 'OR' operation was used for model ensembling, which is suitable for models where the output is categorical (such as the Dtree model). In the future, if the output consists of probabilities for each class, alternative ensemble strategies such as averaging the probabilities, applying a voting mechanism, or selecting the maximum value could be explored, which is a more mathematical approach. Additionally, the number of models in the ensemble can be increased beyond the two considered here. When selecting models for integration, it is important to consider not only the accuracy but also the unique advantages of different models, such as their performance in different spatial areas or periods. This can mitigate the shortcomings of the integrated model and enhance its overall capabilities.

5. Conclusions

In summary, the proposed ensemble model BPNN+Dtree not only increased the hailstorm detection rate but also improved the overall detection accuracy using data from across China through combining the strengths of both models. For other types of hazardous weather besides hail, the risk of missed reports often exceeds that of false alarms, especially in the context of unbalanced samples where such weather events are rare. An integrated learning model can improve the detection rate while ensuring that the false alarm rate does not increase significantly.

Hailstorms, as a mesoscale weather phenomenon, are always considered to be unpredictable. To date, most automatic recognition methods for hailstorms have been developed based on radar echoes, while few take satellite observations into account. The ensemble machine learning model established in this study provides a method for hailstorm detection using satellite data, enabling large-scale automatic monitoring of hail, which contributes to improving the accuracy, timeliness, and refinement level of hail nowcasting and forecasting.

Author Contributions: Conceptualization, Q.W., Y.-X.S., Y.-G.Z., F.W. and C.-Y.W.; Methodology, Q.W. and Y.-X.S.; Validation, Q.W.; Formal analysis, Q.W.; Investigation, Q.W.; Resources, Q.W., Y.-X.S., Y.-G.Z. and F.W.; Data curation, Q.W., Y.-X.S. and Y.-G.Z.; Writing—original draft, Q.W.; Writing—review & editing, Q.W., Y.-X.S., Y.-G.Z. and C.-Y.W.; Visualization, Q.W.; Funding acquisition, Q.W., Y.-X.S. and F.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (grant number 42375140 and 41965001) and the Key Project of the Science and Technology Commission of Shanghai Municipality (grant number 22DZ1100803).

Data Availability Statement: The FY-4A AGRI L1 data were obtained from the National Satellite Meteorological Center (NSMC) of the China Meteorological Administration (CMA), accessible at http://satellite.nsmc.org.cn/PortalSite/Default.aspx (accessed between the years 2018 and 2023). The GPM Precipitation Feature (PF) data were sourced from the Texas A&M University at Corpus Christi (TAMU-CC) precipitation feature database, available at http://atmos.tamucc.edu/trmm/data/ (accessed on 13 May 2024). The GPM Hail Climatology Data Products were retrieved from the NASA Earth Science Data Systems (ESDS) Program, found at https://search.earthdata.nasa.gov/search/granules?p=C2196515446-GHRC_DAAC&pg[0][v]=f&pg[0][gsk]=-start_date&q=hail&tl=1722849455!3!! (accessed on 25 April 2024). The ERA5 reanalysis data were downloaded from the Climate Data Store (CDS), which can be accessed at https://cds.climate.copernicus.eu/cdsapp#!/home (accessed between the years 2020 and 2023). Hail event records were collected from the National Meteoro-

logical Information Center (NMIC) of CMA. Additionally, the elevation data were obtained from Natural Earth, available at https://www.naturalearthdata.com/downloads/10m-natural-earth-1/10m-natural-earth-1-with-shaded-relief-and-water/ (accessed on 9 December 2020).

Acknowledgments: We express our gratitude to all these organizations for their outstanding work in data collection and production.

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

References

1. Hand, W.H.; Cappelluti, G. A global hail climatology using the UK Met Office convection diagnosis procedure (CDP) and model analyses. *Meteorol. Appl.* **2011**, *18*, 446–458. [CrossRef]
2. Zhang, X.; Sun, J.; Zheng, Y.; Zhang, Y.; Ma, R.; Yang, X.; Zhou, K.; Han, X. Progress in Severe Convective Weather Forecasting in China since the 1950s. *J. Meteorol. Res.* **2020**, *34*, 699–719. [CrossRef]
3. Chinese Meteorological Administration. *Yearbook of Meteorological Disasters in China*; China Meteorological Press: Beijing, China, 2013. (In Chinese)
4. Chinese Meteorological Administration. *Yearbook of Meteorological Disasters in China*; China Meteorological Press: Beijing, China, 2014. (In Chinese)
5. Chinese Meteorological Administration. *Yearbook of Meteorological Disasters in China*; China Meteorological Press: Beijing, China, 2015. (In Chinese)
6. Chinese Meteorological Administration. *Yearbook of Meteorological Disasters in China*; China Meteorological Press: Beijing, China, 2016. (In Chinese)
7. Chinese Meteorological Administration. *Yearbook of Meteorological Disasters in China*; China Meteorological Press: Beijing, China, 2017. (In Chinese)
8. Chinese Meteorological Administration. *Yearbook of Meteorological Disasters in China*; China Meteorological Press: Beijing, China, 2018. (In Chinese)
9. Chinese Meteorological Administration. *Yearbook of Meteorological Disasters in China*; China Meteorological Press: Beijing, China, 2019. (In Chinese)
10. Chinese Meteorological Administration. *Yearbook of Meteorological Disasters in China*; China Meteorological Press: Beijing, China, 2020. (In Chinese)
11. Chinese Meteorological Administration. *Yearbook of Meteorological Disasters in China*; China Meteorological Press: Beijing, China, 2021. (In Chinese)
12. Chinese Meteorological Administration. *Yearbook of Meteorological Disasters in China*; China Meteorological Press: Beijing, China, 2022. (In Chinese)
13. Punge, H.J.; Kunz, M. Hail observations and hailstorm characteristics in Europe: A review. *Atmos. Res.* **2016**, *176*, 159–184. [CrossRef]
14. Witt, A.; Eilts, M.D.; Stumpf, G.J.; Johnson, J.T.; Mitchell, E.D.; Thomas, K.W. An enhanced hail detection algorithm for the WSR-88D. *Weather Forecast.* **1998**, *13*, 286–303. [CrossRef]
15. Wang, P.; Shi, J.; Hou, J.; Hu, Y. The Identification of Hail Storms in the Early Stage Using Time Series Analysis. *J. Geophys. Res.-Atmos.* **2018**, *123*, 929–947. [CrossRef]
16. Bang, S.D.; Cecil, D.J. Testing Passive Microwave-Based Hail Retrievals Using GPM DPR Ku-Band Radar. *J. Appl. Meteorol. Climatol.* **2021**, *60*, 255–271. [CrossRef]
17. Cecil, D.J. Passive Microwave Brightness Temperatures as Proxies for Hailstorms. *J. Appl. Meteorol. Climatol.* **2009**, *48*, 1281–1286. [CrossRef]
18. Ni, X.; Liu, C.; Cecil, D.J.; Zhang, Q. On the Detection of Hail Using Satellite Passive Microwave Radiometers and Precipitation Radar. *J. Appl. Meteorol. Climatol.* **2017**, *56*, 2693–2709. [CrossRef]
19. Mroz, K.; Battaglia, A.; Lang, T.J.; Cecil, D.J.; Tanelli, S.; Tridon, F. Hail-Detection Algorithm for the GPM Core Observatory Satellite Sensors. *J. Appl. Meteorol. Climatol.* **2017**, *56*, 1939–1957. [CrossRef]
20. Cecil, D.J.; Blankenship, C.B. Toward a Global Climatology of Severe Hailstorms as Estimated by Satellite Passive Microwave Imagers. *J. Clim.* **2012**, *25*, 687–703. [CrossRef]
21. Bang, S.D.; Cecil, D.J. Constructing a Multifrequency Passive Microwave Hail Retrieval and Climatology in the GPM Domain. *J. Appl. Meteorol. Climatol.* **2019**, *58*, 1889–1904. [CrossRef]
22. Punge, H.J.; Bedka, K.M.; Kunz, M.; Werner, A. A new physically based stochastic event catalog for hail in Europe. *Nat. Hazards* **2014**, *73*, 1625–1645. [CrossRef]
23. Bedka, K.; Brunner, J.; Dworak, R.; Feltz, W.; Otkin, J.; Greenwald, T. Objective Satellite-Based Detection of Overshooting Tops Using Infrared Window Channel Brightness Temperature Gradients. *J. Appl. Meteorol. Climatol.* **2010**, *49*, 181–202. [CrossRef]
24. Bedka, K.M. Overshooting cloud top detections using MSG SEVIRI Infrared brightness temperatures and their relationship to severe weather over Europe. *Atmos. Res.* **2011**, *99*, 175–189. [CrossRef]
25. Adler, R.F.; Markus, M.J.; Fenn, D.D. Detection of severe midwest thunderstorms using geosynchronous satellite data. *Mon. Weather Rev.* **1985**, *113*, 769–781. [CrossRef]

26. Dworak, R.; Bedka, K.; Brunner, J.; Feltz, W. Comparison between GOES-12 Overshooting-Top Detections, WSR-88D Radar Reflectivity, and Severe Storm Reports. *Weather Forecast.* **2012**, *27*, 684–699. [CrossRef]
27. Marion, G.; Trapp, R.J.; Nesbitt, S.W. Using overshooting top area to discriminate potential for large, intense tornadoes. *Geophys. Res. Lett.* **2019**, *46*, 12520–12526. [CrossRef]
28. Punge, H.J.; Bedka, K.M.; Kunz, M.; Reinbold, A. Hail frequency estimation across Europe based on a combination of overshooting top detections and the ERA-INTERIM reanalysis. *Atmos. Res.* **2017**, *198*, 34–43. [CrossRef]
29. Merino, A.; Lopez, L.; Sanchez, J.L.; Garcia-Ortega, E.; Cattani, E.; Levizzani, V. Daytime identification of summer hailstorm cells from MSG data. *Nat. Hazards Earth Syst. Sci.* **2014**, *14*, 1017–1033. [CrossRef]
30. Melcón, P.; Merino, A.; Luis Sanchez, J.; Lopez, L.; Hermida, L. Satellite remote sensing of hailstorms in France. *Atmos. Res.* **2016**, *182*, 221–231. [CrossRef]
31. Bergen, K.J.; Johnson, P.A.; de Hoop, M.V.; Beroza, G.C. Machine learning for data-driven discovery in solid Earth geoscience. *Science* **2019**, *363*, eaau0323. [CrossRef] [PubMed]
32. Bonavita, M.; Arcucci, R.; Carrassi, A.; Dueben, P.; Geer, A.J.; Le Saux, B.; Longepe, N.; Mathieu, P.-P.; Raynaud, L. Machine Learning for Earth System Observation and Prediction. *Bull. Am. Meteorol. Soc.* **2021**, *102*, E710–E716. [CrossRef]
33. Dramsch, J.S. 70 years of machine learning in geoscience in review. *Adv. Geophys.* **2020**, *61*, 1–55. [CrossRef]
34. Fleming, S.W.; Watson, J.R.; Ellenson, A.; Cannon, A.J.; Vesselinov, V.C. Machine learning in Earth and environmental science requires education and research policy reforms. *Nat. Geosci.* **2021**, *14*, 878–880. [CrossRef]
35. Im, J.; Park, H.; Takeuchi, W. Advances in Remote Sensing-Based Disaster Monitoring and Assessment. *Remote Sens.* **2019**, *11*, 2181. [CrossRef]
36. Karpatne, A.; Ebert-Uphoff, I.; Ravela, S.; Babaie, H.A.; Kumar, V. Machine Learning for the Geosciences: Challenges and Opportunities. *Ieee Trans. Knowl. Data Eng.* **2019**, *31*, 1544–1554. [CrossRef]
37. Reichstein, M.; Camps-Valls, G.; Stevens, B.; Jung, M.; Denzler, J.; Carvalhais, N.; Prabhat, F. Deep learning and process understanding for data-driven Earth system science. *Nature* **2019**, *566*, 195–204. [CrossRef] [PubMed]
38. Sehad, M.; Ameur, S. A multilayer perceptron and multiclass support vector machine based high accuracy technique for daily rainfall estimation from MSG SEVIRI data. *Adv. Space Res.* **2020**, *65*, 1250–1262. [CrossRef]
39. Yang, Z.; Liu, P.; Yang, Y. Convective/Stratiform Precipitation Classification Using Ground-Based Doppler Radar Data Based on the K-Nearest Neighbor Algorithm. *Remote Sens.* **2019**, *11*, 2277. [CrossRef]
40. Su, H.; Wu, L.; Jiang, J.H.; Pai, R.; Liu, A.; Zhai, A.J.; Tavallali, P.; DeMaria, M. Applying Satellite Observations of Tropical Cyclone Internal Structures to Rapid Intensification Forecast with Machine Learning. *Geophys. Res. Lett.* **2020**, *47*, e2020GL089102. [CrossRef]
41. Kim, D.; Park, M.-S.; Park, Y.-J.; Kim, W. Geostationary Ocean Color Imager (GOCI) Marine Fog Detection in Combination with Himawari-8 Based on the Decision Tree. *Remote Sens.* **2020**, *12*, 149. [CrossRef]
42. Yao, H.; Li, X.; Pang, H.; Sheng, L.; Wang, W. Application of random forest algorithm in hail forecasting over the Shandong Peninsula. *Atmos. Res.* **2020**, *244*, 105093. [CrossRef]
43. Dotse, S.-Q.; Larbi, I.; Limantol, A.M.; De Silva, L.C. A review of the application of hybrid machine learning models to improve rainfall prediction. *Model. Earth Syst. Environ.* **2024**, *10*, 19–44. [CrossRef]
44. Allen, J.T.; Giammanco, I.M.; Kumjian, M.R.; Punge, H.J.; Zhang, Q.; Groenemeijer, P.; Kunz, M.; Ortega, K. Understanding Hail in the Earth System. *Rev. Geophys.* **2020**, *58*, e2019RG000665. [CrossRef]
45. Hou, A.Y.; Kakar, R.K.; Neeck, S.; Azarbarzin, A.A.; Kummerow, C.D.; Kojima, M.; Oki, R.; Nakamura, K.; Iguchi, T. The global precipitation measurement mission. *Bull. Am. Meteorol. Soc.* **2014**, *95*, 701–722. [CrossRef]
46. Nesbitt, S.W.; Zipser, E.J.; Cecil, D.J. A census of precipitation features in the tropics using TRMM: Radar, ice scattering, and lightning observations. *J. Clim.* **2000**, *13*, 4087–4106. [CrossRef]
47. Liu, C.; Zipser, E.J. The global distribution of largest, deepest, and most intense precipitation systems. *Geophys. Res. Lett.* **2015**, *42*, 3591–3595. [CrossRef]
48. Liu, C.; Zipser, E.J.; Cecil, D.J.; Nesbitt, S.W.; Sherwood, S. A Cloud and Precipitation Feature Database from Nine Years of TRMM Observations. *J. Appl. Meteorol. Climatol.* **2008**, *47*, 2712–2728. [CrossRef]
49. Liu, N.; Liu, C. Global distribution of deep convection reaching tropopause in 1year GPM observations. *J. Geophys. Res.-Atmos.* **2016**, *121*, 3824–3842. [CrossRef]
50. Cecil, D.J.; Bang, S.D. Passive Microwave Hail Climatology Data Products. 2022. Available online: https://doi.org/10.5067/MODEL/DATA101 (accessed on 25 April 2024).
51. Zhou, Z.; Zhang, Q.; Allen, J.T.; Ni, X.; Ng, C.-P. How Many Types of Severe Hailstorm Environments Are There Globally? *Geophys. Res. Lett.* **2021**, *48*, e2021GL095485. [CrossRef]
52. Liu, J.; Ma, C.; Liu, C.; Qin, D.; Gu, X. An extended maxima transform-based region growing algorithm for convective cell detection on satellite images. *Remote Sens. Lett.* **2014**, *5*, 971–980. [CrossRef]
53. Wu, Q.; Shou, Y.-X.; Ma, L.-M.; Lu, Q.; Wang, R. Estimation of Maximum Hail Diameters from FY-4A Satellite Data with a Machine Learning Method. *Remote Sens.* **2022**, *14*, 73. [CrossRef]
54. Kononenko, I.; Simec, E.; RobnikSikonja, M. Overcoming the myopia of inductive learning algorithms with RELIEFF. *Appl. Intell.* **1997**, *7*, 39–55. [CrossRef]

55. Robnik-Sikonja, M.; Kononenko, I. Theoretical and empirical analysis of ReliefF and RReliefF. *Mach. Learn.* **2003**, *53*, 23–69. [CrossRef]
56. Glorot, X.; Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, Sardinia, Italy, 13–15 May 2010; pp. 249–256.
57. Loh, W.Y. Classification and regression trees. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2011**, *1*, 14–23. [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

Quantifying the Geomorphological Susceptibility of the Piping Erosion in Loess Using LiDAR-Derived DEM and Machine Learning Methods

Sisi Li [1,2], Sheng Hu [1,3,4,*] Lin Wang [1,2], Fanyu Zhang [5], Ninglian Wang [1,3,4], Songbai Wu [1,3,4], Xingang Wang [6] and Zongda Jiang [1,2]

[1] Shaanxi Key Laboratory of Earth Surface System and Environmental Carrying Capacity, Northwest University, Xi'an 710127, China; 202322850@stumail.nwu.edu.cn (S.L.); wanglin@nwu.edu.cn (L.W.); nlwang@nwu.edu.cn (N.W.); s.wu@nwu.edu.cn (S.W.); 202233333@stumail.nwu.edu.cn (Z.J.)
[2] School of Information Science and Technology, Northwest University, Xi'an 710127, China
[3] College of Urban and Environmental Sciences, Northwest University, Xi'an 710127, China
[4] Institute of Earth Surface System and Hazards, Northwest University, Xi'an 710127, China
[5] MOE Key Laboratory of Mechanics on Disaster and Environment in Western China, Department of Geological Engineering, Lanzhou University, Lanzhou 730000, China; zhangfy@lzu.edu.cn
[6] State Key Laboratory of Continental Dynamics, Department of Geology, Northwest University, Xi'an 710127, China; xgwang@nwu.edu.cn
* Correspondence: husheng@nwu.edu.cn

Abstract: Soil piping erosion is an underground soil erosion process that is significantly underestimated or overlooked. It can lead to intense soil erosion and trigger surface processes such as landslides, collapses, and channel erosion. Conducting susceptibility mapping is a vital way to identify the potential for soil piping erosion, which is of enormous significance for soil and water conservation as well as geological disaster prevention. This study utilized airborne radar drones to survey and map 1194 sinkholes in Sunjiacha basin, Huining County, on the Loess Plateau in Northwest China. We identified seventeen key hydrogeomorphological factors that influence sinkhole susceptibility and used six machine learning models—support vector machine (SVM), logistic regression (LR), Convolutional Neural Network (CNN), K-Nearest Neighbors (KNN), random forest (RF), and gradient boosting decision tree (GBDT)—for the susceptibility assessment and mapping of loess sinkholes. We then evaluated and validated the prediction results of various models using the area under curve (AUC) of the Receiver Operating Characteristic Curve (ROC). The results showed that all six of these machine learning algorithms had an AUC of more than 0.85. The GBDT model had the best predictive accuracy (AUC = 0.94) and model migration performance (AUC = 0.93), and it could find sinkholes with high and very high susceptibility levels in loess areas. This suggests that the GBDT model is well suited for the fine-scale susceptibility mapping of sinkholes in loess regions.

Keywords: loess sinkholes; piping; LiDAR; machine learning; susceptibility mapping

1. Introduction

In karst areas, vertical cracks widen due to water erosion, or collapses occur at the tops of underground rivers, often forming surface shapes resembling funnel-shaped sinkholes that act as significant pathways for surface runoff to enter underground [1,2]. Apart from karst regions, various climate zones and soil types worldwide exhibit similar landforms [3–6]. Due to differences in their formation environments and causes, they are generally referred to as pseudokarst landforms [7–9]. Especially in arid and semi-arid environments, the largest and most typical sinkholes develop in loess or derived soils [3,10,11]. In the Loess Plateau of China, a large number of loess sinkholes are also widely distributed and developing [11]. This is mainly attributed to factors such as high sand content, low

organic matter content, a loose structure, high porosity, well-developed vertical joints, strong self-weight settlement, low erosion resistance, easy disintegration, the development of preferential flow paths, and serious dry–wet cycles induced by long-term drought and waterlogging alternation [4,12]. Soil pipes are not only a phenomenon of underground soil erosion but also a geological hazard. Many reports indicate that these sinkholes often lead to road collapses, house collapses, farmland destruction, dam pipeline blockages, and railway damage. Under certain conditions, sinkholes not only have a significant impact on the stability of slopes but can also trigger other geological disasters such as mudslides and landslides [4,5,13,14]. Therefore, conducting surveys and the mapping of sinkholes is of great significance for the prevention and control of geological disasters and soil erosion.

In recent decades, some countries have begun to realize the importance of this work. Traditional sinkhole investigations and mapping primarily rely on topographic maps, digital elevation models (DEMs), historical aerial photography, or satellite images [2,15–18]. In recent years, many researchers have started using remote sensing [19], synthetic aperture radar interferometry (InSAR) [20], photogrammetry [21], and light detection and ranging (LiDAR) [22–26] to detect and identify sinkholes worldwide. For instance, high-resolution remote sensing satellite images and unmanned aerial system (UAS) images can be used for interpreting, cataloging, and mapping sinkholes [19]; InSAR technology can monitor the rate of sinkhole formation or ground subsidence [20]; and airborne LiDAR measurements can be employed for the deep detection of large sinkholes and terrain mapping [23,24]. An increasing amount of high-resolution topographic data is being utilized to identify and detect sinkholes [27–30]. Due to rugged and densely vegetated terrain, traditional survey methods struggle to achieve a detailed surveying and mapping [28,31]. Notably, LiDAR technology has uncovered more sinkholes in various karst regions [32–34]. For instance, using LiDAR data, Zhu [34] discovered three times as many sinkholes compared to those previously identified on topographic maps in the Floyds Fork watershed of Central Kentucky. Additionally, researchers have autonomously extracted karst collapse features based on high-resolution LiDAR DEMs, local profile tree methods, and weighted random forest approaches [35,36]. Currently, most sinkhole detection is carried out in karst environments, where sinkhole diameters and depths usually range from tens to hundreds of meters, making them easy to detect. In contrast, the sinkholes formed in loess regions tend to be smaller, usually measuring several to tens of meters in both diameter and depth [3,4]. The efficacy of advanced technologies in surveying such areas of the Loess Plateau has yet to be validated through field investigations. It is well-established that the Loess Plateau in China boasts the widest and deepest accumulation of loess in the world, resulting in numerous loess sinkholes [11,37]. Research in this domain remains in its infancy [38], particularly regarding the application of advanced laser scanning technologies for foundational surveys and assessments of loess caves, which have been reported infrequently. Therefore, we selected a typical study area in the Northwest Loess Plateau of China to investigate loess sinkholes using an airborne radar UAS.

Machine learning is a branch of artificial intelligence that constructs computer-based systems that automatically improve through training data [39]. Its main advantage is the ability to analyze the complex input–output relationships of physical behaviors using measured data and handle data space patterns on various scales [40]. In recent years, various machine learning models have increasingly been applied to the automatic recognition, assessment, and mapping of sinkholes [27,29,36,41,42]. For instance, Verachtert [43] utilized an LR model to predict the spatial patterns of sinkholes in loess soils within a moist temperate climate. They observed that sinkholes emerge when topographic thresholds dependent on slope and contributing area are met, and that in well-drained regions, an excessive slope and contributing area led to them. Hosseinalizadeh [44] utilized three machine learning algorithms—MDA, flexible discriminant analysis (FDA), SVM, and UAS imagery—to map the susceptibility of pipeline erosion in the loess-covered hilly regions of Northeastern Iran, demonstrating that all three algorithms effectively predict sinkhole susceptibility. Taheri and his colleagues used various Bayesian-based machine learning

models to predict the spatial distribution of sinkholes in Iran, selecting the information gain ratio technique based on the regulating factors of sinkholes [41]. On the other hand, Peyman employed SVM, artificial neural network (ANN), and decision tree (DT) models using soil characteristics obtained from Central Iran and ground-penetrating radar (GPR) frequency data to model and forecast the hazards of sinkholes [45]. The current studies mostly use two to three machine learning models to assess sinkhole susceptibility, with little comparison to recent models that have shown better evaluation results within the same spatial framework. This research employed six classic machine learning methods—SVM, LR, CNN, KNN, RF, and GBDT—for the susceptibility modeling and spatial mapping of sinkholes in the study area.

Because there are many factors affecting the development of sinkholes, many scholars have utilized various types of data, such as meteorological, hydrological, topographic, geological, soil type, land use, and soil properties, for their susceptibility assessment [46–50]. However, for small catchments, the data on soil, geology, lithology, and rainfall may not have much spatial heterogeneity on the one hand, and there may be a lack of high-resolution data or a difficulty in obtaining such data on the other hand, all of which restrict the susceptibility mapping of sinkholes. In fact, hydrogeomorphologic processes at a small watershed scale play a crucial role in the development of sinkholes. Therefore, this paper tries to use an easily obtained DEM and its derived hydrogeomorphic factors to quickly predict the sinkhole susceptibility of small watersheds.

2. Study Area

The study area is located in Huining County, on the Loess Plateau in Northwest China. During the Early Jurassic–Mesozoic era, this area was once a vast lake basin. Due to the rapid uplift of the Qinghai–Tibet Plateau and mountain-building movements, the ancient lake basin was uplifted into loess hills and exposed on the surface. Since the Late Pleistocene, strong northwest winds have continuously blown sand and dust particles from the deserts and desert regions of Northwest China and Central Asia to this area, gradually depositing them here and forming the embryonic form of loess hills [11]. The regional geological map (1:200,000) shows that most of the slopes in the study area are covered by Q_3 wind-blown loess (Malan loess), while flood or alluvial layers formed by Q_4 are mainly distributed in the valley areas. The region is characterized by numerous gullies due to long-term rainfall and water erosion, low vegetation cover, severe soil erosion, a relatively fragile geological environment, intense human engineering activities, and prominent environmental geological issues of geological hazards and soil erosion in the research area.

The selected key investigation area in this study is the Sunjiacha basin in Huining County (Figure 1a,b), with an area of about 2.40 km^2, a length of approximately 2960 m, a width of around 1280 m, and an elevation ranging from 1724.10 to 2070.88 m. It has a typical combination of loess ridges and gullies, which are part of the semi-arid temperate monsoon climate. The surface vegetation mainly consists of sparse grasslands, with precipitation concentrated from May to September, totaling about 370 mm annually. Rainfall from a single heavy rain event can account for 96% of the monthly precipitation. The river channel normally experiences no runoff except during heavy or torrential rain periods, when runoff is generated [11]. The small watershed is characterized by numerous loess geological hazards (1194 collapsing sinkholes, 288 landslides, and collapses) with deep channels and severe soil erosion (mainly by gully erosion). Preliminary field surveys and studies revealed a high sinkhole density of up to 497.5 per km^2, with a maximum depth of 29.6 m, a maximum diameter of 34 m, and a maximum volume of approximately 19,600 m^3. The study area distributes these sinkholes in various locations, including gullies, valley heads, valley terraces, and terraced fields. The imagery and topography obtained from the airborne radar drone in the research area are shown in Figure 1c,d.

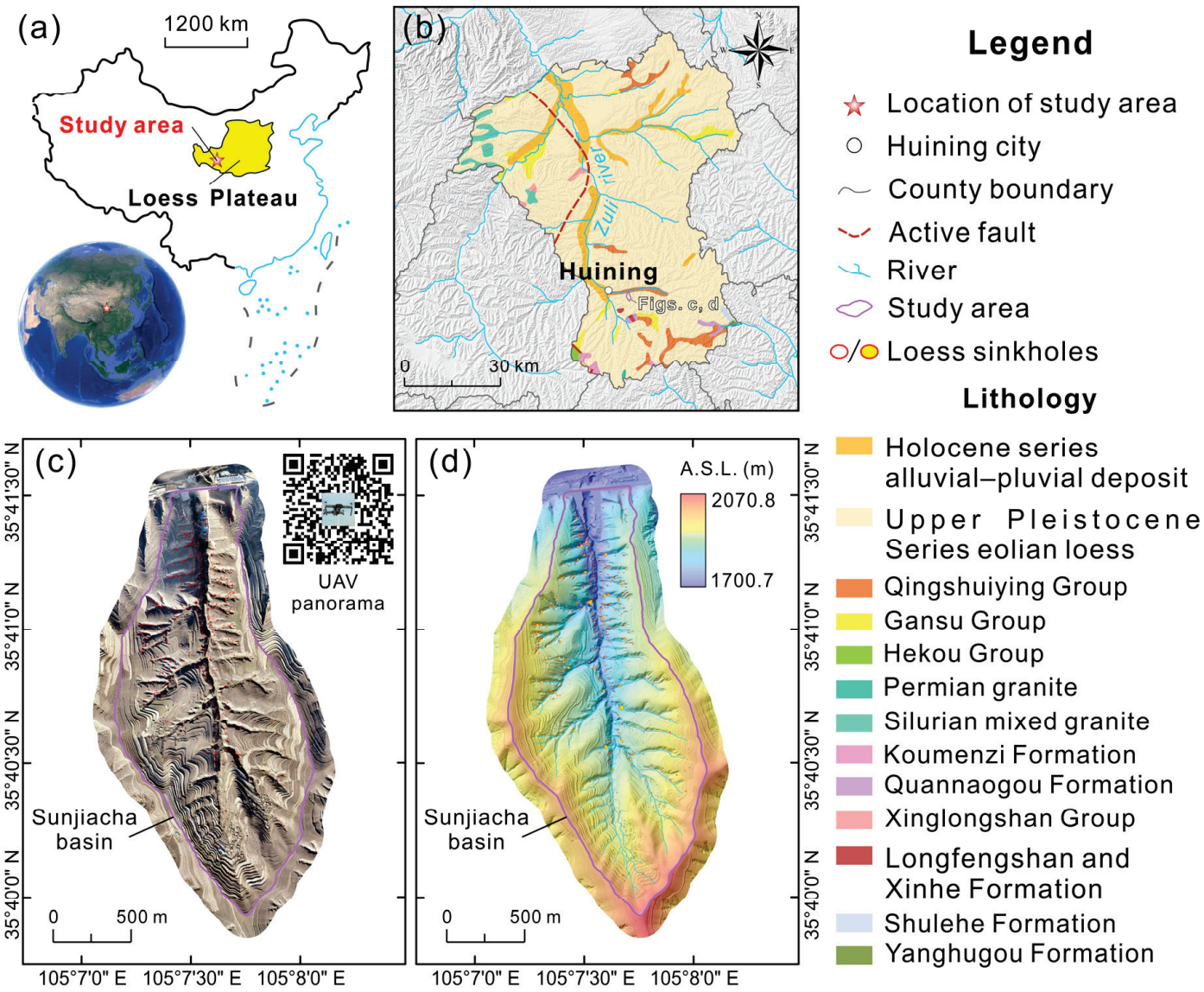


Figure 1. Study area overview: (**a**) location; (**b**) regional geological map; (**c**) digital orthophoto map (DOM) by UAS optical camera; (**d**) LiDAR-derived DEM.

3. Materials and Methods

The technical process of this study, as shown in Figure 2, mainly consists of five steps: (i) using the UAS to collect digital orthophoto images and original LiDAR point cloud data of the research area; (ii) preprocessing the data by figuring out how to interpret UAS images to find sinkholes, cleaning up and filtering point clouds to make DEMs, and making terrain factors; (iii) building a model to test the susceptibility of sinkholes and training it with six machine learning models; (iv) mapping the susceptibility of sinkholes using different machine learning models; and (v) confirming and choosing the best model.

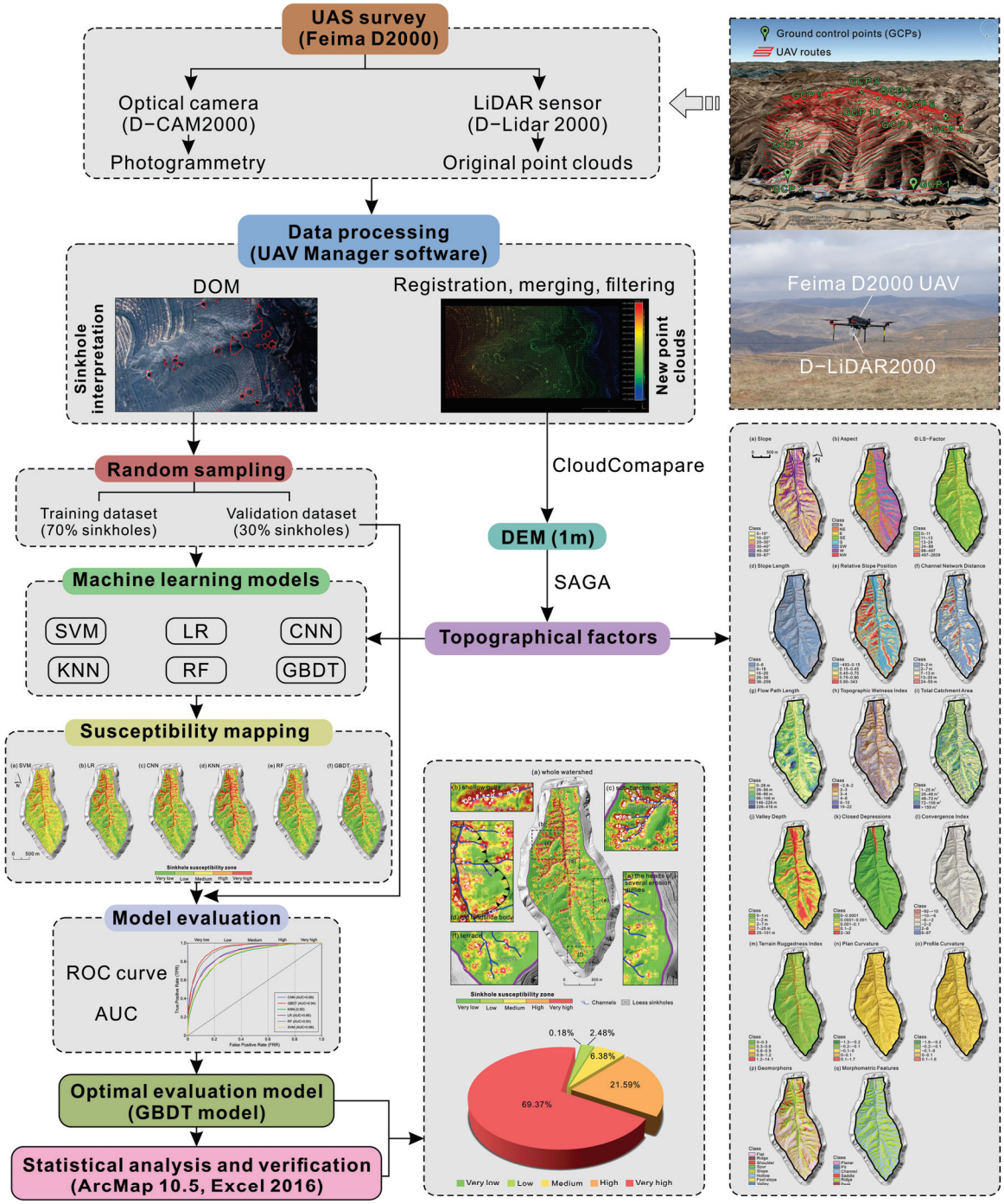


Figure 2. The technical flow chart of this study.

3.1. UAS Survey

After identifying the research area, we conducted on-site field investigations to study and further interpret the sinkholes' morphology. To perform the two flight missions, this study utilized the Feima D2000 unmanned aerial vehicle system equipped with the D-Lidar 2000 laser radar and D-CAM2000 optical camera. For detailed technical parameters of the drone, please visit http://www.feimarobotics.com/zhcn/productDetailD2000 (accessed on 15 April 2024). Point cloud data and imagery of the research area were obtained during the process (Figure 3a–c). During the UAS flight planning, the flight height was set at 200 m, with a longitudinal overlap rate of 70% and a lateral overlap rate of 60%. This flight path design ensured both image overlap and resolution, thereby enhancing surveying accuracy

and quality. The D-Lidar 2000 laser radar features triple-echo technology, enabling excellent vegetation penetration and more comprehensive terrain point cloud data acquisition. To ensure the precision of the drone measurements, we positioned 10 ground control points (GCPs) within and around the watershed and utilized a handheld real-time kinematic (RTK) device to measure these GCPs (Figure 3d–g). Following the field surveys, the data preprocessing of the collected photos and point cloud data was conducted in UAS Manager Professional 1.6.7, resulting in the production of high-resolution digital orthophoto maps (DOM) and raw point cloud data (with an average density of 192 points/m^2). Figure 3h,i depict the point cloud and DOM of the same local area, while Figure 3j–l showcase on-site investigation photos of the sinkholes in the study area.

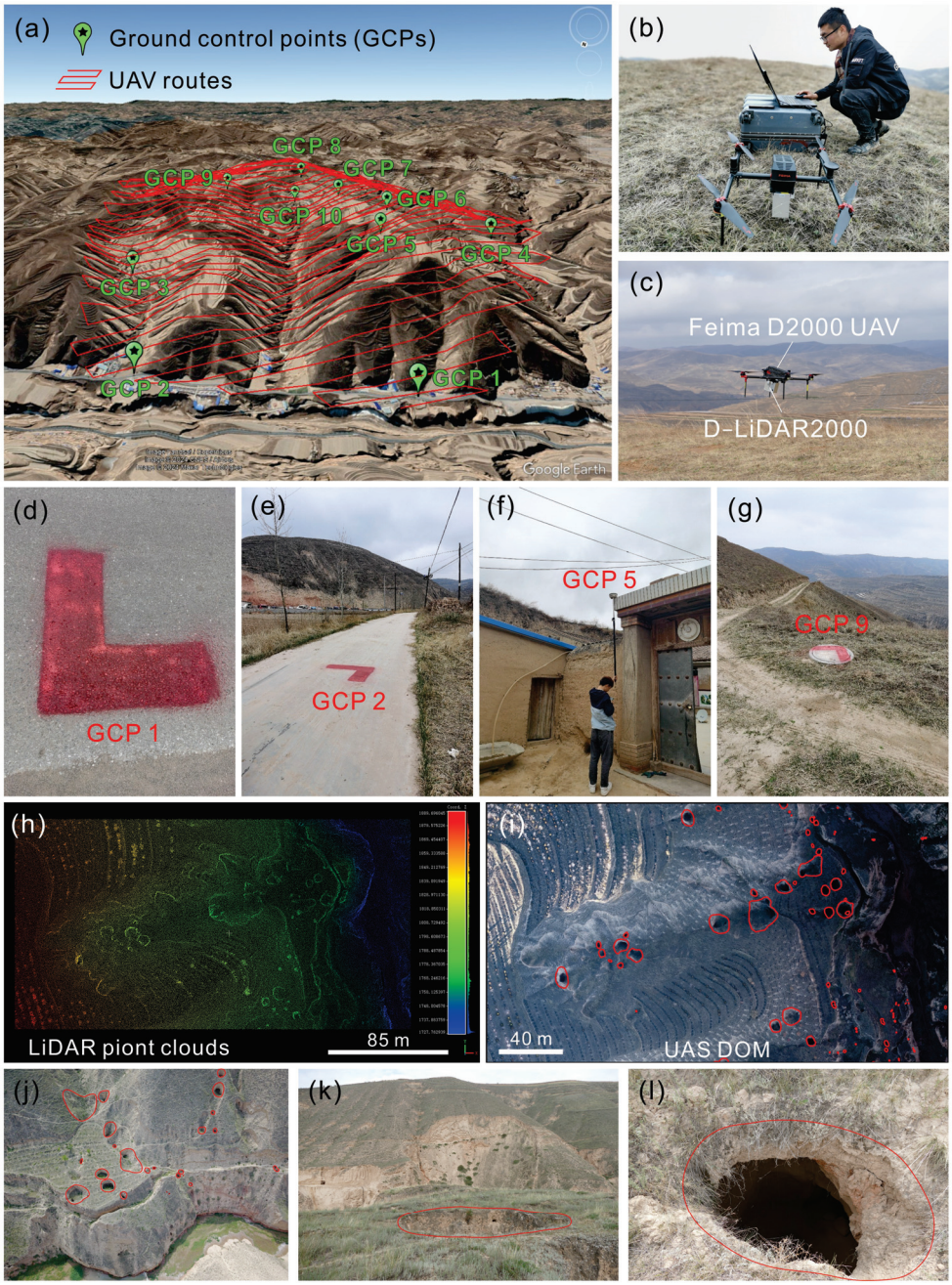


Figure 3. UAS survey and results: (**a**) route planning; (**b**,**c**) Feima D2000 UAS; (**d**–**g**) ground control point surveying using handheld RTK; (**h**) local point clouds acquired by D-LiDAR2000; (**i**) local DOM acquired by D-CAM2000; (**j**–**l**) typical sinkhole photos taken in field. The red circle is the artificial interpretation of the sinkhole polygon in (**i**–**l**).

3.2. Data Preparation

3.2.1. Sinkholes Inventory

By modeling the photos captured by the UAS optical camera, a drone DOM with a spatial resolution of 6.87 cm was obtained. Subsequently, 1194 sinkholes were interpreted in ArcGIS 10.5, extracting the key morphological parameters of the sinkholes from the acquired point cloud, DEM, and DOM data, including major axis, minor axis, elongation ratio, roundness, perimeter, area, depth, and volume. Among the 1194 sinkholes, parameters such as major axis, minor axis, elongation ratio, roundness, perimeter, and area were successfully extracted, while 809 sinkholes had depth and volume parameters successfully extracted. Sinkholes with a length of less than 1 m were disregarded due to their extremely small size, resulting in insufficient point cloud data acquisition during aerial photography, making depth and volume calculations unfeasible. The interpreted sinkholes ranged from a maximum area of 662.18 m^2 to a minimum of 0.03 m^2, with an average size of 17.75 m^2. The deepest detected sinkhole had a depth of 29.60 m, while the shallowest was 0.42 m, with an average depth of 6.55 m. The statistical analysis of these morphological parameters is presented in Table 1.

Table 1. Main morphometric parameters calculated from the sinkhole inventory.

Parameter	Maximum	Minimum	Average	Standard Deviation	Sample Quantity
* Major axis (m)	34.68	0.19	3.78	4.09	1194
* Minor axis (m)	25.12	0.15	2.83	3.10	1194
* Elongation ratio	7.35	1.00	1.41	0.42	1194
* Circularity ratio	0.98	0.33	0.89	0.08	1194
* Perimeter (m)	104.14	0.60	10.75	11.74	1194
* Area (m^2)	662.18	0.03	17.75	46.25	1194
** Depth (m)	29.60	0.42	6.55	4.51	809
** Volume (m^3)	19601.27	0.21	334.75	1119.00	809

* represents the parameters extracted from the UAS-DOM; ** represents the parameters extracted and calculated from the LiDAR data. Volume = Area × Depth.

3.2.2. LiDAR Point Clouds Generate DEM

We processed the 40 GB of raw point cloud data obtained from the preprocessing in CloudCompare V2, resulting in a more complete terrain point cloud that filtered out noise such as buildings, people, cars, vegetation, towers, and power lines, and generated a 1 m resolution DEM from this terrain point cloud data (Figure 1d).

3.2.3. Dataset Preparation

We loaded the obtained DEM raster data into ArcMap 10.5 and resampled the sinkhole and non-sinkhole areas separately. Specifically, we conducted random sampling in the sinkhole areas by creating random points, and uniform sampling in the non-sinkhole areas using a fishnet grid with 20 m intervals. In the end, we obtained 10,880 sampling points for sinkhole areas and 8762 sampling points for non-sinkhole areas, labeled "1" and "0", respectively. After preparing the dataset, we divided the 19,642 samples into two subsets for training and testing using a random selection method. We used around 70% of the sampling points (13,749) for training and 30% (5893) for testing.

3.3. Selection of Geomorphic Factors

After obtaining the sinkhole dataset, another crucial step involved identifying the susceptibility factors that influence sinkhole development. However, there is currently no unified guideline for selecting parameters for a sinkhole susceptibility assessment [51,52]. Baker and Shatz [53,54] indicate that the selection process for geological environmental factors is driven by the principle of parsimony. This principle emphasizes the importance of using the minimum necessary elements to effectively explain phenomena. Therefore, through field investigations and a relevant literature analysis [47,49,55,56], we ultimately

selected 17 topographic factors as key indicators for assessing the susceptibility of loess sinkholes. These factors included slope, aspect, slope length factor (LS-factor), slope length, relative slope position, vertical distance to the channel network, flow path length, topographic wetness index, total catchment area, valley depth, closed depressions, convergence index, terrain ruggedness index, plan curvature, profile curvature, geomorphons, and morphometric features. Among these, slope, aspect, slope length factor, slope length, topographic wetness index, convergence index, terrain ruggedness index, plan curvature, and profile curvature are commonly used topographic and hydrological factors for assessing sinkhole susceptibility [44,57]. Additionally, we incorporated the geological environmental conditions of the study area and introduced factors that suit regions with minimal spatial variability, such as basin closure, landform type, morphological measurement characteristics, and valley depth. Table 2 presents the descriptions and sources of the geomorphic factors. We calculated the 17 topographic factors based on a LiDAR-derived DEM using the open-source System for Automated Geoscientific Analyses (SAGA, https://saga-gis.sourceforge.io/en/index.html (accessed on 15 April 2024)) GIS (Version 7.7.0) and completed the mapping in ArcMap 10.5 (Figure 4). All data layers are in raster format with a spatial resolution of 1 m.

Table 2. Hydrological and geomorphic factors used in the study.

Variable	Data Type	Scale	Description	References
Slope	Raster	1 m	Calculate topographic parameters slope	Florinsky [58]
Aspect	Raster	1 m	The orientation of the topography or land surface	Florinsky [58]
LS-Factor	Raster	1 m	The slope length factor in the soil erosion equation (USLE) is calculated based on the slope and (specific) watershed area	Kinnell [59]
Slope Length	Raster	1 m	Calculate the slope (or flow path) length	Florinsky [58]
Relative Slope Position	Raster	1 m	The topographic index related to topographic position is calculated from a digital elevation model using an iterative method	Boehner and Selige [60]
Channel Network Distance	Raster	1 m	Used to calculate the vertical distance to a channel network base level	Passalacqua et al. [61]
Flow Path Length	Raster	1 m	The average flow path length starting from the selected grid cells	Quinn et al. [62]
Topographic Wetness Index	Raster	1 m	Utilized to assess the propensity of an area to accumulate water based on its terrain characteristics	Boehner and Selige [60]
Total Catchment Area	Raster	1 m	Refine the calculation using the flow accumulation tool	Gruber and Peckham [63]
Valley Depth	Raster	1 m	Difference between the elevation and an interpolated ridge level	Skentos [64]
Closed Depressions	Raster	1 m	A depression created by the erosion or dissolution of the earth's surface	Kołodyńska-Gawrysiak et al. [65]
Convergence Index	Raster	1 m	Calculates an index of convergence/divergence regarding overland flow	Krawchuk et al. [66]
Terrain Ruggedness Index	Raster	1 m	Used to quantify terrain ruggedness and topographic complexity	Riley et al. [67]
Plan Curvature	Raster	1 m	The curvature of the Earth's surface as observed in a horizontal plane	Heerdegen and Beran [68]
Profile Curvature	Raster	1 m	The curvature of the Earth's surface as observed along a line or profile, typically in the direction of maximum slope	Heerdegen and Beran [68]
Geomorphons	Raster	1 m	Derived from a digital elevation model using machine vision approaches	Jasiewicz and Stepinski [69]
Morphometric Features	Raster	1 m	A multi-scale approach is used to derive data for morphological feature classification by fitting quadratic parameters	Wood [70]

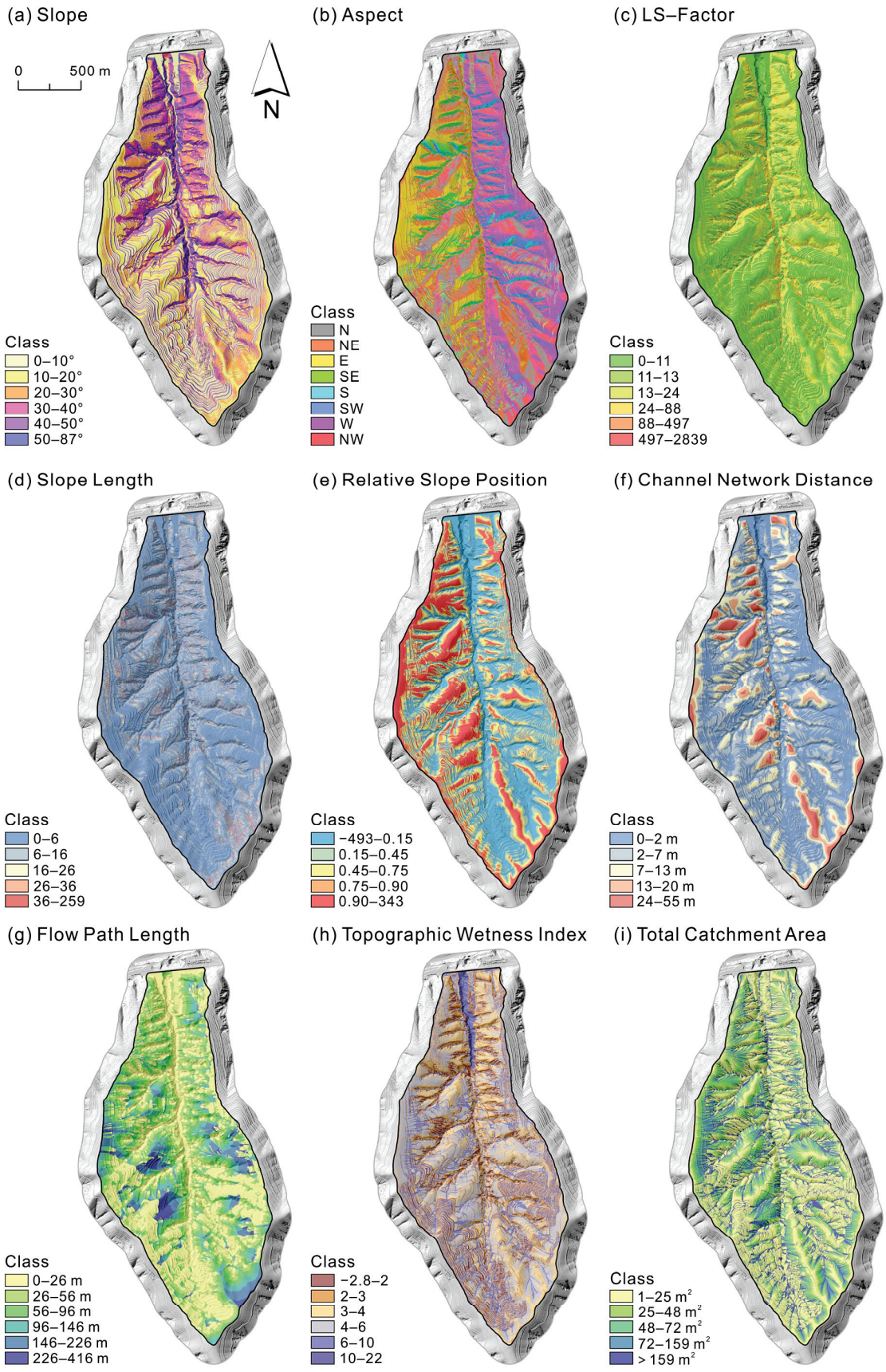


Figure 4. *Cont.*

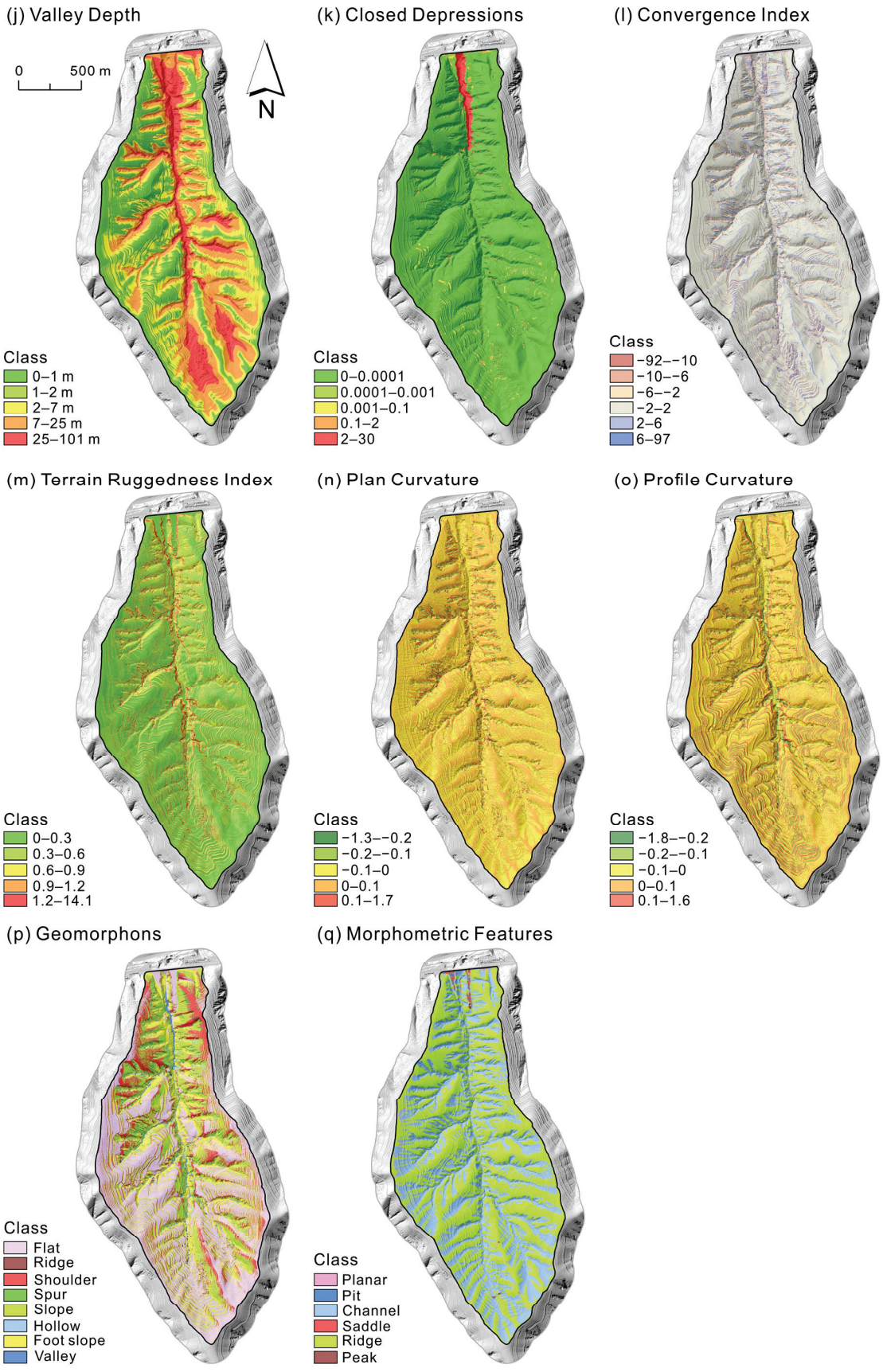


Figure 4. Geomorphic factor mapping for evaluating sinkhole susceptibility.

3.4. Machine Learning Models

When it comes to susceptibility assessment in sinkhole risk analysis, the SVM model is the best statistical and machine learning method for making predictions [71,72]. SVM is a machine learning method based on statistical learning theory, applicable to classification and regression analysis [73]. Its key strength lies in transforming the initial input space into a high-dimensional feature space, estimating an optimal separating hyperplane, and accurately classifying new unknown samples into known categories [74,75]. This method works well for complicated issues involving both linear and nonlinear separations. It uses different kernel functions, such as linear, polynomial, sigmoid activation, and radial basis functions, to separate samples efficiently [76]. In this study, we selected the radial basis kernel function, whose performance can be optimized by adjusting the kernel width [77]. By finely tuning the kernel width parameter within the range of 1 to 10, we can better control the model's complexity and generalization ability. Additionally, the regularization parameter influences the accuracy of the SVM model, with a reasonable selection within the range of 0.01 to 1.0 aiding in balancing the model's fitting capability and preventing overfitting. We used the grid search method to look for the best combination of parameters within the given parameter range. We chose the combination that yielded the highest accuracy score in the validation data to construct the SVM model.

In this study, LR, as a multivariate statistical method, aims to construct predictive models of probabilities for specific categories or events, suitable for binary response scenarios such as yes or no, success or failure, and other classification issues. This method plays a significant role in creating susceptibility maps for geological hazards, for instance in assessing the risk of floods [78] and sinkholes [79]. In our research, we consider the presence of sinkholes as the dependent variable, and we predict sinkholes using logistic regression, which offers significant advantages over traditional linear regression models that require the dependent variable to follow a normal distribution [79,80]. Given the complex formation mechanism of sinkholes, which frequently does not conform to normal distribution characteristics, LR models better accommodate the nature of such non-normally distributed data, resulting in more accurate and reliable results in assessing geological hazard risks. Therefore, this study selects logistic regression as the primary analytical tool to effectively evaluate the risk of sinkhole disasters.

A CNN is a type of feedforward neural network architecture based on deep learning widely used in the field of visual image analysis [81]. It is characterized by the use of convolution operations in at least one network layer, replacing traditional matrix multiplication. A typical CNN consists of an input layer, an output layer, and multiple hidden layers, with the hidden layers primarily composed of a series of convolutional layers. These convolutional layers extract features from input data using convolutional operations that are distinct from conventional multiplication or dot product operations. Following the convolutional layers, there are often other types of layers, such as pooling layers, fully connected layers, and normalization layers, collectively forming the hidden layer section. Influenced by activation functions and the final convolution operation, the inputs and outputs of these layers are termed hidden layers [82]. In the training process of the CNN in this study, as the evaluation stage progresses, in order to reduce computational workload and enhance feature invariance, it is necessary to decrease the spatial resolution of the data to reduce the storage capacity of spatial information. Additionally, pooling functions are introduced to reduce the data dimensions while retaining essential feature information. Furthermore, the final convolution layers widely apply the backpropagation algorithm to optimize the network parameters and improve feature extraction accuracy by more precisely adjusting the network weights. This unique network structure and learning mechanism enable the CNN to achieve a high-precision feature extraction and classification performance in visual image analysis.

The KNN algorithm in the classification process does not determine the category of a new instance through an explicit learning process; rather, it relies on the category patterns of its k nearest neighbors to make a decision. This characteristic makes KNN an

instance-based learning algorithm that classifies by directly comparing training data [83]. In the KNN model, the selection of the k value, distance measurement, and classification rules are three core elements. During classification, the k value determines how many neighboring pieces of information are considered, and its magnitude significantly affects the classification results. We use distance measurement to calculate the similarity between a new instance and the various instances in the training dataset, using common distance measurement methods such as Euclidean distance and Manhattan distance. In this study, we chose Euclidean distance as the distance measurement method because it yields the minimal distance [84]:

$$L_{(x_i,x_j)} = \left(\sum_{i=1,j=1}^{n} \left|x_i - x_j\right|^2\right)^{1/2} \quad (1)$$

In this study, x_i and x_j represent the i and j feature vectors. For the classification rule, the majority voting rule was adopted, which is based on determining the category of a new instance according to the distribution of k nearest neighbors. This classification rule is characterized by objectivity and universality, making it suitable for classification problems in most scenarios.

Random forest, as an ensemble learning method, aims to construct multiple weak classification trees using the bootstrap technique and aggregate the results of these trees to classify unknown samples [85]. In this learning process, each iteration involves selecting predictive variables and reordering data with replacements [86,87]. Through this process, the RF model demonstrates a better ability to avoid overfitting issues and generally exhibits an improved generalization performance. The performance of a random forest is influenced by two key structural parameters. The first parameter is the number of features used in each random tree (mtry); by adjusting this parameter, the exploration of the feature space by each tree can be controlled, thereby affecting the model's complexity and classification performance. The second parameter is the number of trees in the random forest (ntree); increasing the number of trees usually helps to enhance the model's stability and accuracy, but it may also increase the computational complexity and training time [88,89]. In our case, the optimal values for mtry and ntree were estimated using the grid search technique [74].

GBDT is an iterative decision tree algorithm that focuses on utilizing the boosting ensemble method. This method was introduced by Friedman [90] to address overfitting issues in decision trees. GBDT uses regression trees as the basic classifiers and employs the gradient descent method in each iteration to build new basic classifiers. Each basic classifier learns the gradient descent direction of the loss function from the previous iteration. When the GBDT model converges, its final prediction result is the weighted sum of all the base classifiers' outputs. It is noteworthy that, unlike in RF, where each classifier has the same weight, in GBDT, the classifiers have different weights, reflecting their varied contributions to the final prediction result. GBDT demonstrates superior performance compared to RF in solving certain classification problems, although it has limitations such as the inability to parallel train and a high susceptibility to data noise. We built the GBDT model in this study using the Python software package version 2.7.9.

3.5. Validation and Accuracy Assessment

In the evaluation of geological hazard susceptibility, previous researchers often used the ROC to validate the zoning results, evaluating the accuracy of zoning results by calculating the AUC value [91]. The confusion matrix has two axes that show the actual values and the predicted values. The *TP* (true positive) axis shows the correct prediction of sinkholes, the *TN* (true negative) axis shows the correct prediction of non-sinkholes, the *FP* (false positive) axis shows the misclassification of non-sinkholes as sinkholes, and the *FN* (false negative) axis shows the misclassification of sinkholes as non-sinkholes. We can calculate the overall accuracy, which is the ratio of all the correctly predicted points to the

total number of points, and obtain the true positive rate (TP_{Rate}) and false positive rate ($FP_{\text{Rate)}}$ from the confusion matrix:

$$TP_{\text{Rate}} = \frac{TP}{TP + FN} \tag{2}$$

$$FP_{\text{Rate}} = \frac{FP}{FP + TN} \tag{3}$$

The TP_{Rate} represents the ratio of correctly predicted sinkhole points to the total number of sinkhole points, while the FP_{Rate} represents the ratio of points incorrectly predicted as non-sinkholes to the total number of non-sinkhole points. When making a comparison using the area under the ROC curve, when AUC $\leq$ 0.5, it indicates that the predictive ability of the evaluation model is equivalent to or even lower than random guessing, thus the model lacks predictive value; when 0.5 < AUC $\leq$ 0.7, it indicates a certain level of accuracy in the predictive ability of the evaluation model; when 0.7 < AUC $\leq$ 0.9, it indicates a good accuracy in the predictive ability of the evaluation model; and when AUC $\geq$ 0.9, it indicates the high predictive ability of the evaluation model.

4. Results

4.1. Susceptibility Mapping of Different Machine Learning Models

Based on the machine learning models (SVM, LR, CNN, KNN, RF, and GBDT) and 17 terrain factors, we created a sinkhole susceptibility map (Figure 5). We divided each susceptibility map into five susceptibility levels: very low, low, medium, high, and very high. Based on these six susceptibility maps, we generated corresponding frequency distribution histograms. Among these, GBDT closely approximates the actual scenario with an exponential decay frequency distribution histogram. SVM, which has the lowest prediction accuracy, shows a probability distribution graph that approximates a normal distribution and differs significantly from the actual scenario. LR and CNN have very similar probability distribution graphs (resembling the Poisson distribution), and their susceptibility mapping is highly comparable. The probability distribution graphs of KNN and RF closely resemble the actual scenario, although with some differences: the KNN model has higher values in the 0.6–1 range, while RF has lower values in the 0–0.1 range.

Figure 6 shows the actual areas for each susceptibility level of the six models, with the SVM model indicating that low susceptibility has the largest area (0.93 km^2), followed by medium susceptibility (0.87 km^2), high susceptibility (0.49 km^2), lower susceptibility (0.09 km^2), and higher susceptibility (0.01 km^2). Based on the RF model, the area of lower susceptibility is the largest (0.82 km^2), followed by low susceptibility (0.68 km^2), medium susceptibility (0.44 km^2), high susceptibility (0.30 km^2), and higher susceptibility (0.15 km^2). In the CNN model, the largest area is in the lower susceptibility category at 0.72 km^2, followed by 0.70 km^2 (low susceptibility), 0.46 km^2 (medium susceptibility), 0.33 km^2 (high susceptibility), and 0.18 km^2 (higher susceptibility). In the GBDT model, lower susceptibility has the largest area (1.05 km^2), followed by low susceptibility (0.52 km^2), medium susceptibility (0.35 km^2), high susceptibility (0.28 km^2), and higher susceptibility (0.19 km^2). According to the LR model, the largest area of 0.73 km^2 is for the low susceptibility category, followed by 0.70 km^2 (lower susceptibility), 0.46 km^2 (medium susceptibility), 0.33 km^2 (high susceptibility), and 0.19 km^2 (higher susceptibility). In the KNN model, the largest area (for lower susceptibility) is 0.76 km^2, followed by low susceptibility (0.52 km^2), medium susceptibility (0.38 km^2), higher susceptibility (0.37 km^2), and high susceptibility (0.37 km^2). The general trend of area variation from the six models is the same as the frequency distribution histograms. This means that the areas with low and medium susceptibility are generally bigger in each model, while the areas with medium, high, and high susceptibility get smaller over time.

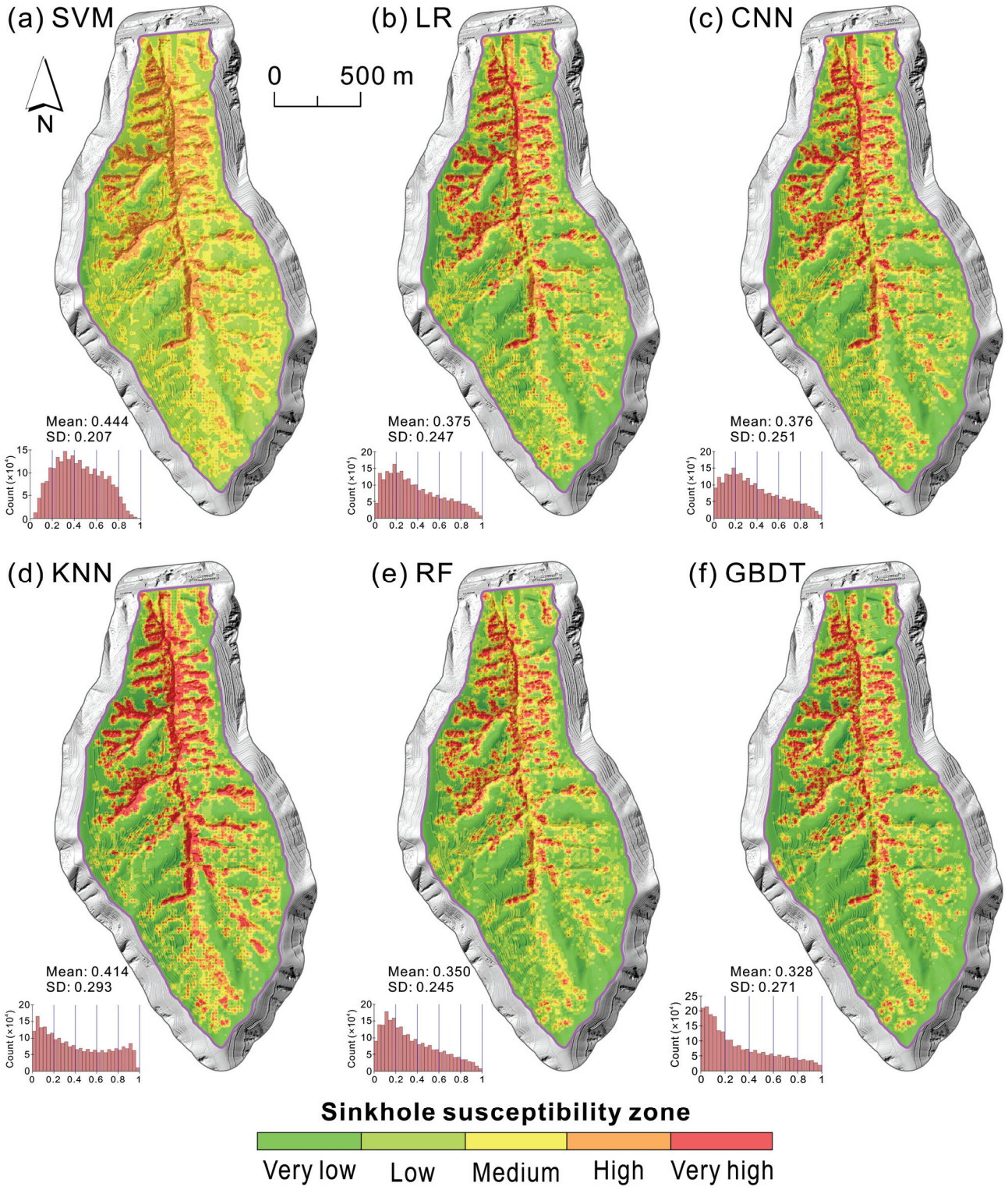


Figure 5. Sinkhole susceptibility maps and frequency distribution histograms of grid by six machine learning methods, (**a**) SVM, (**b**) LR, (**c**) CNN, (**d**) KNN, (**e**) RF, and (**f**) GBDT, where mean stands for average and SD stands for standard deviation.

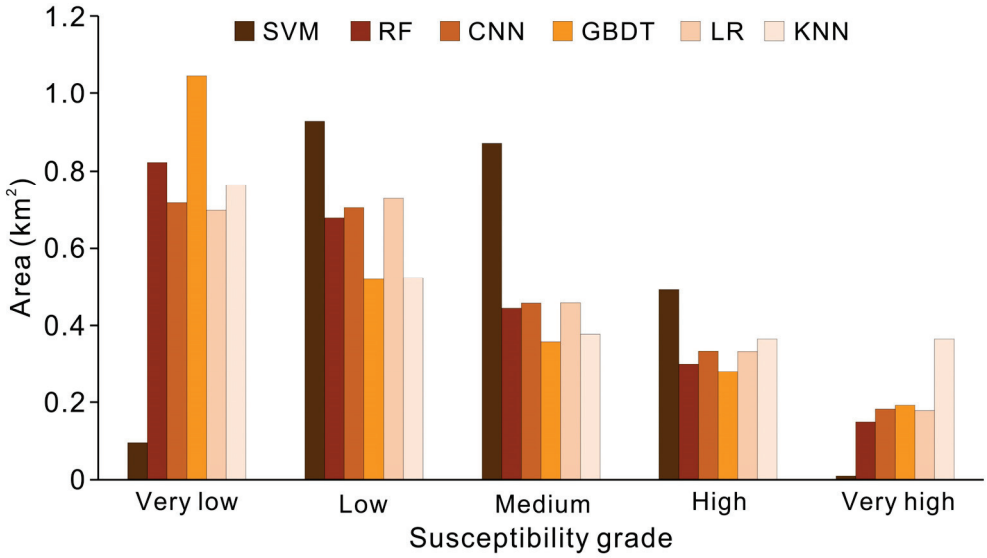


Figure 6. Areal comparison of susceptibility grades for six models.

4.2. Verification Comparison of Models in Susceptibility Prediction

As shown in Figure 7, the validation results indicate that the GBDT model has the highest AUC value (0.94), followed by RF with 0.93, KNN with 0.90, CNN with 0.89, LR with 0.88, and SVM with 0.86. Accordingly, this study concludes that the predictive abilities of the GBDT, RF, and KNN models are excellent, with AUC values $\geq$ 0.9, while the CNN, LR, and SVM models also demonstrate a good predictive accuracy, with AUC values $\geq$ 0.8. The AUC values of GBDT and RF differ only by 0.01. However, susceptibility mapping shows that GBDT is more precise in identifying small and densely distributed sinkholes. Despite having a high AUC value, susceptibility mapping reveals a poor predictive performance for KNN. This is because the model uses a majority voting rule to predict whether a pixel is a sinkhole, resulting in high capability to distinguish sinkholes from non-sinkholes but a low accuracy in classifying susceptibility levels into five categories. Compared to other machine learning models, the CNN model requires a large amount of watershed data for effective training. However, our collection of watershed data is not as extensive as the terrain factor data, resulting in the CNN model's relatively low predictive capability. Despite significant differences in feature extraction and training methods between the LR and CNN models, the AUC value of the LR model differs by only 0.01 from the CNN model. The susceptibility mapping and frequency distribution histograms of both models are very similar. The SVM model essentially distinguishes different categories based on decision boundaries. In this study, the complex data structure and issues like noise significantly affect the predictive performance, resulting in a lower prediction accuracy compared to the other five models.

The principle of GBDT involves iteratively training decision trees to minimize the loss function. Each tree attempts to fit the residuals of the predictions made by all previous trees, thereby gradually enhancing the model performance. The predictions made by the GBDT model rely on the aggregation of predictions from all trees, making it particularly sensitive to outliers. Notably, the GBDT and RF models exhibit the highest prediction accuracy in this study, as both are widely used algorithms in ensemble learning that employ multiple decision trees for decision-making. Our research focuses on a smaller dataset that encompasses numerous terrain-influencing factors. There exist complex nonlinear relationships among these factors. Tree-based ensemble learning algorithms, such as GBDT and RF, excel at capturing and managing these intricate relationships. Furthermore, they are less prone to overfitting on small datasets and can achieve a better generalization performance compared to deep learning methods.

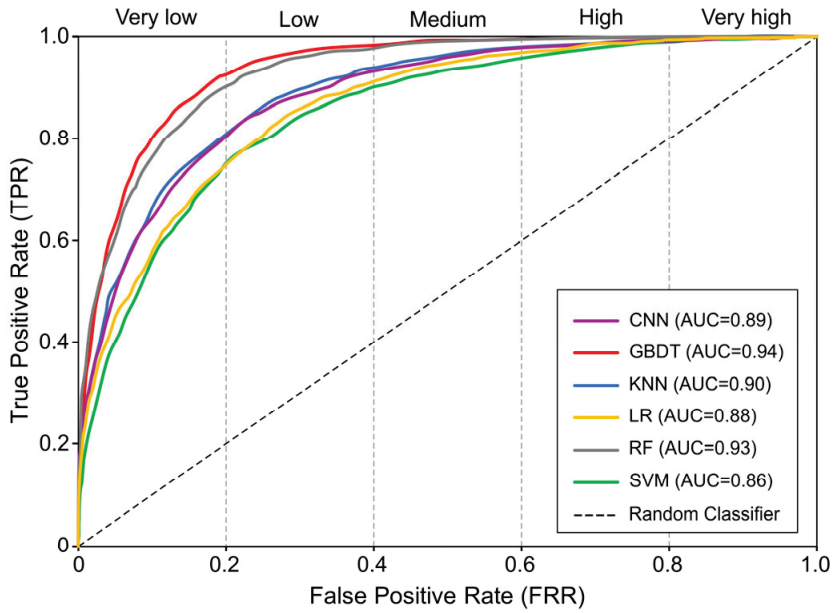


Figure 7. Comparison of the area under the ROC curves (AUC) for six models in the validation step.

4.3. Analysis and Verification of Optimal Susceptibility Mapping

By comparing various assessments, we observed that the GBDT model exhibits the highest predictive accuracy. This model not only performs exceptionally well overall but also demonstrates strong predictive capabilities in identifying sinkholes within different geomorphic units of the study area (Figure 8a). For instance, the lower part of a watershed is where surface water collects, and during heavy rainfall, it often generates surface runoff. In this area, the soil moisture is high, and with water flow erosion, it easily develops bead-like sinkholes (Figure 8b,c). Due to loosening soil and developed fractures, old landslide bodies are susceptible to developing sinkholes under the influence of rainfall and overland flow (Figure 8d). Transitional zones between positive and negative terrain, where localized depression areas are prone to sinkhole formation, are home to headward erosion gullies. Sinkhole incision often causes headward erosion in these regions, which are potential areas for sinkhole development (Figure 8e). The edges of terraced fields develop small sinkholes due to surface runoff erosion (Figure 8f). It is worth mentioning that the GBDT model performs very well in predicting the distribution of sinkholes in areas with dense slopes and channel sinkholes, as well as small, hard-to-identify sinkholes present in old landslides, eroded gully heads, and terraced fields. Through extensive data training, the GBDT model can accurately capture the influence of terrain factors on the distribution of sinkholes and accurately identify areas of dense sinkhole distribution.

Figure 9 shows the proportion of sinkhole areas calculated by the GBDT model at different susceptibility levels. In areas with very high susceptibility, the proportion of sinkhole areas is 69.37%, while the other proportions are 21.59% (high susceptibility), 6.38% (medium susceptibility), 2.48% (low susceptibility), and 0.18% (lower susceptibility), respectively. The total of medium, low, and lower susceptibility does not exceed 10%, while the proportion of sinkhole areas with very high and high susceptibility reaches 90.97%. The GBDT model's susceptibility assessment in the study area meets the expected standards, predicting loess sinkholes locations as primarily falling into the categories of higher and high susceptibility. Overall, the advantages of the GBDT model are more pronounced in predicting these areas compared to the other methods.

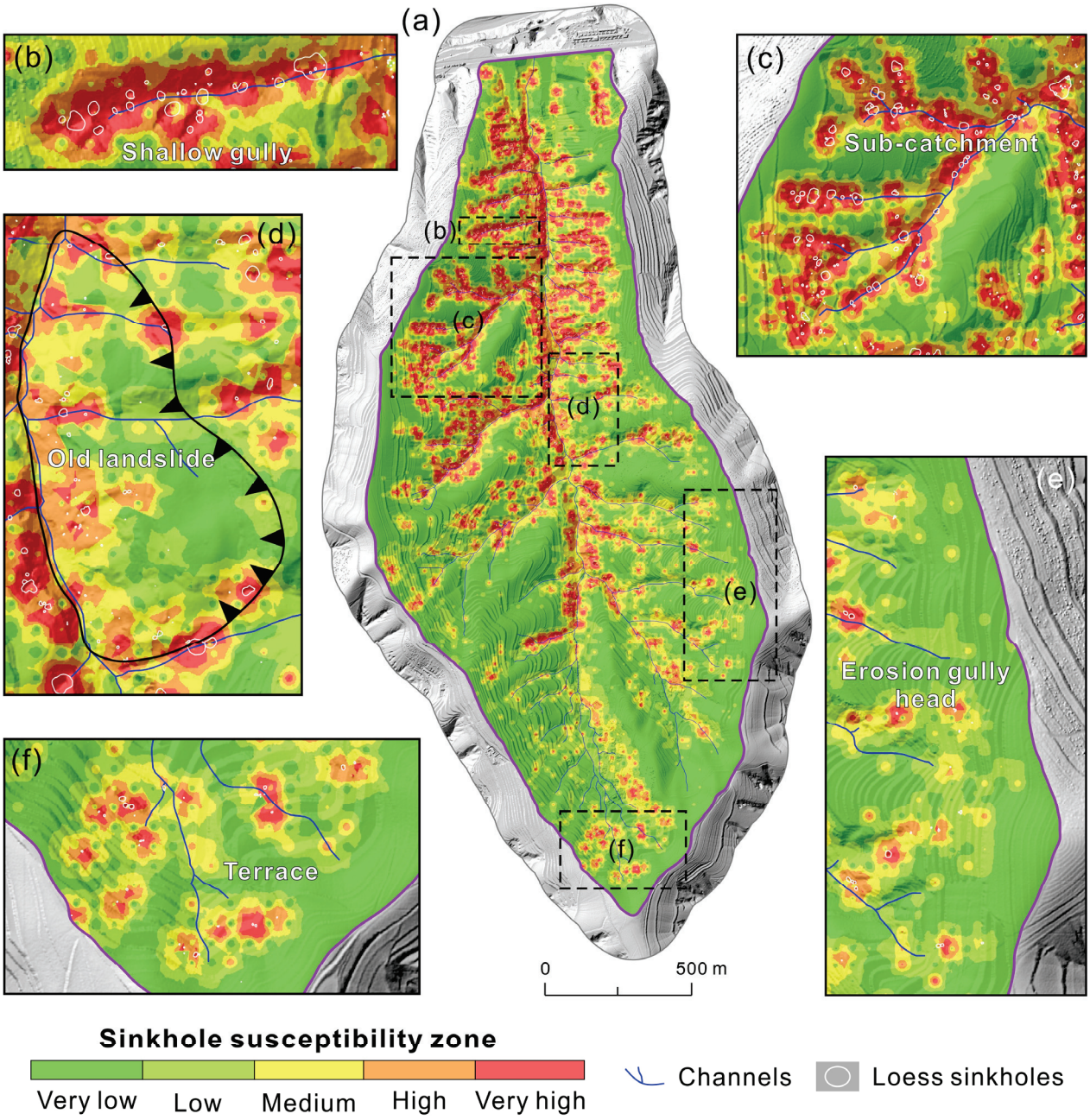


Figure 8. Geomorphological susceptibility mapping of loess sinkholes based on the GBDT model: (**a**) whole watershed; (**b**) shallow gully; (**c**) sub-catchment; (**d**) old landslide body; (**e**) the heads of several erosion gullies; (**f**) terrace.

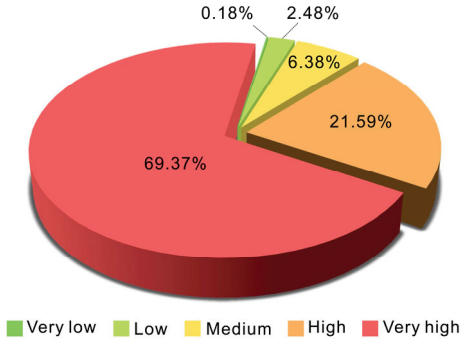


Figure 9. The sinkhole area proportion of different susceptibility grades by the GBDT model.

4.4. Evaluation of Model Migration

To further evaluate the transferability of the six machine learning models employed in this study, we conducted a verification at a loess tableland in Beiguo Village, Heyang County, Shaanxi Province (110.2520°E, 35.0248°N). We mapped 779 loess sinkholes in this area through drone imagery. During the model validation phase, we utilized the same hydrological and geomorphological factors as outlined in Table 2. The validation results indicated (Figure 10) that all six machine learning models achieved AUC values exceeding 0.85, with the gradient boosting decision trees (GBDT) model displaying the highest expected accuracy (AUC = 0.93). This suggests that the GBDT model maintains an excellent transfer capability and high prediction accuracy for sinkholes even when applied to new regions. Additionally, the RF, CNN, LR, SVM, and KNN models also demonstrated commendable transferability. We illustrate the GBDT model's predictions of the geomorphological susceptibility of loess sinkholes in Figure 11. The areas identified as high and very high susceptibility primarily coincide with the tops of old landslides, the landslide body, and erosional gullies, closely matching the actual distribution of sinkholes.

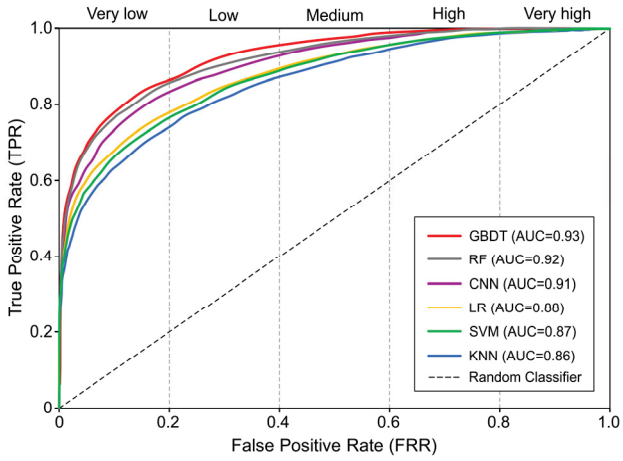


Figure 10. Comparison of area under ROC curve (AUC) of six models in validation area.

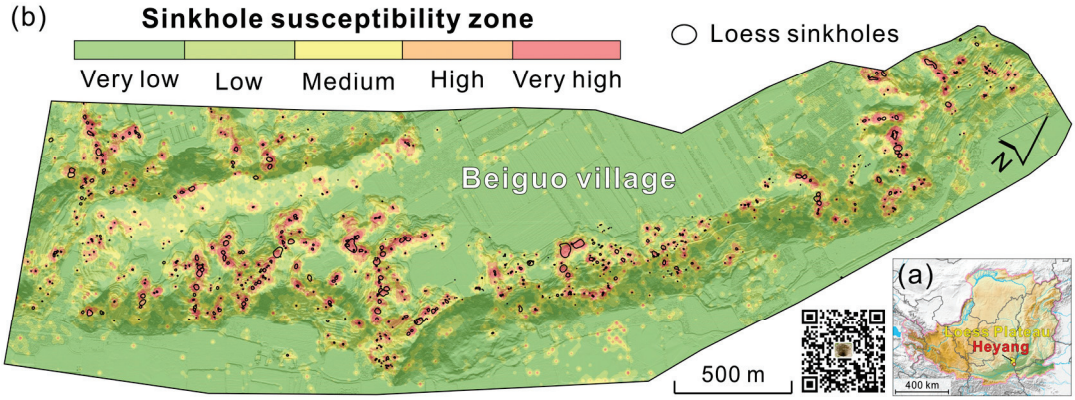


Figure 11. Geomorphological susceptibility mapping by GBDT model in validation area: (**a**) location of validation area; (**b**) susceptibility zoning map by GBDT model.

5. Discussion

5.1. Model Comparison

Over the past two decades, advancements in measurement technology and equipment have enabled many countries to establish a comprehensive detection and monitoring system for sinkholes, spanning from space to low altitudes, near-surface, and underground. However, each technology possesses unique advantages and limitations. For instance, some researchers [21,92–94] have successfully detected sinkholes using unmanned aerial

systems (UASs) or Helikite balloons equipped with various sensors such as optical lenses, LiDAR, and thermal cameras. This technology allows for extensive aerial surveys but has a limited capability for detecting structures deep underground. Additionally, methods combining geomorphological mapping, InSAR, ERT, GPR, drilling, and trenching for sinkhole identification [20,95,96] significantly enhance data accuracy but demand high environmental and operational standards, resulting in relatively high costs. Currently, most sinkhole detection occurs in karst environments, where sinkholes can be quite large. In contrast, sinkholes formed in China's Loess Plateau typically have smaller diameters and depths. The effectiveness of the aforementioned non-destructive testing technologies and methods still requires validation through field investigations. The Loess Plateau in China features complex soil conditions, numerous ravines, and a fragmented topography, leading to the frequent occurrence of landslides and debris flows. This makes the Loess Plateau a unique geological environment [97–99]. Loess sinkholes represent a specific geological hazard in this region, yet research on these areas remains in its infancy [38]. Therefore, we selected a typical small watershed in the Northwest Loess Plateau of China as our study area, aiming to utilize the most commonly employed airborne radar UAS in karst environments to survey and investigate loess sinkholes.

Sensitivity modeling and mapping sinkholes have become essential tools for assessing karst features. We can categorize the sensitivity models of sinkholes into qualitative and quantitative types [100]. Qualitative methods often rely on the expertise and experience of local specialists, such as the Analytic Hierarchy Process (AHP), which may introduce subjective bias [101,102]. In contrast, quantitative methods depend on numerical data and a statistical analysis to interpret phenomena, employing techniques such as statistics, multi-criteria decision-making, and machine learning [101,103,104]. Recently, advancements in machine learning have significantly improved its predictive capabilities, offering more accurate and efficient solutions [105,106]. Consequently, many researchers have begun to utilize data from various sensors and develop algorithms to swiftly extract and analyze sinkhole information. For instance, Hosseinalizadeh [44] predicted sinkhole sensitivity in the loess-covered hilly areas of Northeastern Iran using a drone system and three machine learning algorithms: mixed discriminant analysis (MDA), flexible discriminant analysis (FDA), and support vector machine (SVM). The validation results indicated that the AUC for these three algorithms ranged from 92.45% to 90.32%. Zhu [42] successfully identified 97% of sinkholes in the Bluegrass region of Kentucky by utilizing morphological feature data from LiDAR and applying six machine learning methods—logistic regression, naive Bayes, neural networks, random forests, RUSBoost, and support vector machines—with neural networks performing the best. However, their study relied solely on morphological feature data from sinkholes without incorporating influencing factors. In contrast, this research combines the morphological feature data of sinkholes with 17 terrain factors to identify sinkholes in the study area and classify their disaster sensitivity. While neural networks excel at handling spatial data, their advantages may be surpassed by the ensemble learning strategy of random forests when dealing with more diverse and complex feature sets, such as terrain factors. Additionally, support vector machines and logistic regression exhibit weaker capabilities in managing complex and high-dimensional feature sets. Specifically, SVM requires more computational resources to find the optimal hyperplane as features increase and become more complex, potentially leading to performance degradation. Chen [107] investigated the predictive performance of four machine learning models (a boosted linear model, boosted regression trees, boosted generalized linear model, and deep boosting model) based on 18 factors for mapping pipeline erosion sensitivity in the Zarandieh watershed of Markazi Province, Iran. The deep boosting model achieved an area under the ROC curve of 0.93, demonstrating its capability in mapping pipeline erosion sensitivity. This study employs another classic algorithm of the boosting ensemble method, gradient boosting decision trees (GBDT), achieving an AUC value of 0.94. Notably, the deep boosting model often requires substantial data for training and may encounter issues such as overfitting and complex parameter tuning. In contrast,

GBDT tends to train relatively quickly, especially with medium-sized datasets, and experiments indicate that GBDT outperforms the deep boosting model in predicting sinkhole sensitivity. Encouragingly, compared to sinkhole sensitivity mapping results from other regions [42,44,104,107], this study achieves more refined results by selecting key terrain factors for sinkhole sensitivity mapping.

5.2. Assumptions and Limitations

In karst regions, numerous scholars have conducted extensive and productive research. They have made significant advancements in areas such as the formation processes and causes of sinkholes [56], morphological feature analysis [108,109], automatic extraction and identification [36], and sensitivity mapping [44,107]. However, research on sinkholes in loess areas remains scarce. Although some researchers have progressed in classifying them, understanding their distribution patterns, and identifying disaster causes related to loess sinkholes [4,5,11,50,110], overall, these studies are still in their infancy [38]. Therefore, future in-depth research on the automatic identification, sensitivity assessment, morphological characteristics, and formation mechanisms of sinkholes in loess regions holds significant practical value for preventing and managing soil erosion and geological disasters in these areas.

We employed six machine learning algorithms, utilizing data from 1194 sinkholes and 17 terrain factors, to investigate the disaster sensitivity of the Sunjiacha basin sinkholes in Huining County on the Loess Plateau in Northwest China. The results were satisfactory. All the models demonstrated strong predictive capabilities. We also conducted transfer experiments in new watersheds; the models met expectations, indicating excellent transferability. Notably, the gradient boosting decision tree model accurately identified sinkholes with high and very high sensitivity levels, enabling a detailed analysis of sinkhole sensitivity.

However, it is undeniable that our current work has certain limitations. First, the influencing factors used for training are derived from high-resolution DEM data. For larger watersheds, relying solely on topographic data may not fully capture the complex conditions that lead to the formation of sinkholes. This limitation can hinder the generalization ability of the model and restrict the learning capacity and predictive accuracy of the machine learning model. In the future, we will continue to explore and incorporate additional factors, such as soil property, precipitation, and land use, and test and improve the model over a broader area. Second, in terms of identification results, a small number of smaller sinkholes were not detected due to limited data, while some surface depressions closely resembled sinkholes, leading to misclassification by the algorithm. In future work, we will further optimize or enhance the machine learning model to improve the accuracy of sinkhole identification. Nevertheless, this research demonstrates the potential and value of using LiDAR and machine learning models for predicting sinkhole disaster sensitivity in loess environments. It also provides new insights for the sensitivity assessment and mapping of sinkholes in other loess regions, thereby offering a fresh perspective for the early identification, monitoring, and risk assessment of geological hazards.

6. Conclusions

This study utilized UAS LiDAR measurement technology to conduct a detailed investigation of sinkholes in the Sunjiacha basin in Huining County, Northwest Loess Plateau. We obtained and interpreted high-precision, high-resolution images, terrain data, and LiDAR point cloud data to identify 1194 sinkholes. The sinkhole data were resampled, with 70% used as a training dataset and 30% as a test dataset. Six machine learning operational environments were established on the server. The training dataset and 17 terrain factors were inputted into the machine learning models for training, resulting in susceptibility maps for the six models. Sinkhole susceptibility maps were evaluated using the AUC. Our research indicates the following:

(1) High-resolution LiDAR measurement technology effectively penetrates low, sparse vegetation to directly acquire a precise surface elevation, making it suitable for sink-

hole investigations. This technology is great at fixing the big errors that come with regular UAS image-based terrain reconstructions. It shows a lot of promise for use in investigating sinkholes;

(2) In the evaluation phase, the AUC values of the six machine learning models were 0.94 (GBDT), 0.93 (RF), 0.9 (KNN), 0.89 (CNN), 0.88 (LR), and 0.86 (SVM). Among these, the GBDT, RF, and KNN models demonstrated excellent predictive capabilities, with AUC values $\geq$ 0.9, while the CNN, LR, and SVM models also showed a good accuracy in their predictions, with AUC values $\geq$ 0.8. This indicates that all the models can effectively predict the susceptibility map of sinkholes;

(3) The GBDT model has the best performance in predicting the susceptibility of loess sinkholes in Huining County (AUC = 0.94), especially in predicting potential sinkholes at the head of erosion gullies, the interior of erosion gullies, old landslides, and terraces;

(4) The results of the transferability evaluation of the machine learning models indicate that the six models employed in this study exhibit a good adaptability in new regions (AUC > 0.85). Among all models, the GBDT model shows a superior transfer capability and predictive accuracy (AUC = 0.93).

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

Funding: This study was funded by the National Key Research and Development Program of China (Grant No. 2023YFC3008401), the National Natural Science Foundation of China (Grant Nos. 42371009, 42001006), the Undergraduate Talent Training Construction Project of Northwest University (Grant No. 2023-38), and the Northwest University Graduate Innovation Project (Grant No. CX2024212).

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Kaufmann, G.; Romanov, D. Structure and evolution of collapse sinkholes: Combined interpretation from physico-chemical modelling and geophysical field work. *J. Hydrol.* **2016**, *540*, 688–698. [CrossRef]
2. Gökkaya, E.; Gutiérrez, F.; Ferk, M.; Görüm, T. Sinkhole development in the Sivas gypsum karst, Turkey. *Geomorphology* **2021**, *386*, 107746. [CrossRef]
3. Bernatek-Jakiel, A.; Poesen, J. Subsurface erosion by soil piping: Significance and research needs. *Earth-Sci. Rev.* **2018**, *185*, 1107–1128. [CrossRef]
4. Peng, J.; Sun, P.; Igwe, O. Loess caves, a special kind of geo-hazard on loess plateau, northwestern China. *Eng. Geol.* **2018**, *236*, 79–88. [CrossRef]
5. Hu, S.; Qiu, H.; Wang, N.; Wang, X.; Ma, S.; Yang, D.; Wei, N.; Liu, Z.; Shen, Y.; Cao, M.; et al. Movement process, geomorphological changes, and influencing factors of a reactivated loess landslide on the right bank of the middle of the Yellow River, China. *Landslides* **2022**, *19*, 1265–1295. [CrossRef]
6. Kariminejad, N.; Sepehr, A.; Poesen, J.; Hassanli, A. Combining UAV remote sensing and pedological analyses to better understand soil piping erosion. *Geoderma* **2023**, *429*, 116267. [CrossRef]
7. Chen, C.K. The loess landform types and their development rules in the southeast of Longdong area. *Acta Geogr. Sin.* **1956**, *22*, 223–231. (In Chinese)
8. Halliday, W.R. Pseudokarst in the 21st century. *J. Cave Karst Stud.* **2007**, *69*, 103–113.
9. Asanidze, L.; Lezhava, Z.; Tsikarishvili, K.; Gaprindashvili, G.; Chikhradze, N. Karst and pseudokarst landscape of Georgia (Caucasus). A short review. *Int. Multidiscip. Sci. GeoConf. SGEM* **2021**, *21*, 59–64.
10. Hosseinalizadeh, M.; Kariminejad, N.; Campetella, G.; Jalalifard, A.; Alinejad, M. Spatial point pattern analysis of piping erosion in loess-derived soils in Golestan Province, Iran. *Geoderma* **2018**, *328*, 20–29. [CrossRef]

11. Hu, S.; Qiu, H.; Wang, N.; Cui, Y.; Wang, J.; Wang, X.; Ma, S.Y.; Yang, D.D.; Cao, M. The influence of loess cave development upon landslides and geomorphologic evolution: A case study from the northwest Loess Plateau, China. *Geomorphology* **2020**, *359*, 107167. [CrossRef]
12. Wang, L.; Li, X.A.; Li, L.C.; Hong, B.; Liu, J. Experimental study on the physical modeling of loess tunnel-erosion rate. *Bull. Eng. Geol. Environ.* **2019**, *78*, 5827–5840. [CrossRef]
13. Verachtert, E.; Van Den Eeckhaut, M.; Poesen, J.; Deckers, J. Spatial interaction between collapsed pipes and landslides in hilly regions with loess-derived soils. *Earth Surf. Process. Landf.* **2013**, *38*, 826–835. [CrossRef]
14. Bernatek-Jakiel, A.; Nadal-Romero, E. Can soil piping impact environment and society? Identifying new research gaps. *Earth Surf. Process. Landf.* **2023**, *48*, 72–86. [CrossRef]
15. Panno, S.V.; Weibel, C.P.; Li, W. Karst Regions of Illinois. Illinois State Geological Survey. Open File Series No. 1997-2. 1997. Available online: https://www.ideals.illinois.edu/items/44958/bitstreams/133464/data.pdf (accessed on 25 August 2024).
16. Luman, D.E.; Panno, S.V. Mapping palimpsest karst features on the Illinois sinkhole plain using historical aerial photography. *Carbonate Evaporite* **2013**, *28*, 201–214.
17. De Carvalho Júnior, O.A.; Guimarães, R.F.; Montgomery, D.R.; Gillespie, A.R.; Gomes RA, T.; de Souza Martins, É.; Silva, N.C. Karst depression detection using ASTER, ALOS/PRISM and SRTM-derived digital elevation models in the Bambuí Group, Brazil. *Remote Sens.* **2014**, *6*, 330–351. [CrossRef]
18. Vajedian, S.; Motagh, M. Extracting sinkhole features from time-series of TerraSAR-X/TanDEM-X data. *ISPRS J. Photogramm. Remote Sens.* **2019**, *150*, 274–284. [CrossRef]
19. Dou, J.; Li, X.; Yunus, A.P.; Paudel, U.; Chang, K.T.; Zhu, Z.; Pourghasemi, H.R. Automatic detection of sinkhole collapses at finer resolutions using a multi-component remote sensing approach. *Nat. Hazards* **2015**, *78*, 1021–1044. [CrossRef]
20. Talib, O.C.; Shimon, W.; Sarah, K.; Tonian, R. Detection of sinkhole activity in West-Central Florida using InSAR time series observations. *Remote Sens. Environ.* **2022**, *269*, 112793. [CrossRef]
21. Al-Halbouni, D.; Holohan, E.P.; Saberi, L.; Alrshdan, H.; Sawarieh, A.; Closson, D.; Walter, T.R.; Dahm, T. Sinkholes, subsidence and subrosion on the eastern shore of the Dead Sea as revealed by a close-range photogrammetric survey. *Geomorphology* **2017**, *285*, 305–324. [CrossRef]
22. Filin, S.; Baruch, A. Detection of sinkhole hazards using airborne laser scanning data. *Photogramm. Eng. Remote Sens.* **2010**, *76*, 577–587. [CrossRef]
23. Launspach, J. Automated Sinkhole Extraction and Morphological Analysis in Northeast Iowa Using High-Resolution LiDAR Data. Master's Thesis, University of Northern Iowa, Cedar Falls, IA, USA, 2013.
24. Kobal, M.; Bertoncelj, I.; Pirotti, F.; Dakskobler, I.; Kutnar, L. Using lidar data to analyse sinkhole characteristics relevant for understory vegetation under forest cover—Case study of a high karst area in the Dinaric Mountains. *PLoS ONE* **2015**, *10*, e0122070. [CrossRef] [PubMed]
25. Zhang, S.; Bogus, S.; Baros, S.; Neville, P.; Dow, R. Karst sinkhole detecting and mapping using airborne LiDAR-A conceptual framework. In *MATEC Web of Conferences*; EDP Sciences: Les Ulis, France, 2019; Volume 271, p. 02005.
26. Zhang, S.; Bogus, S.M.; Baros, S.V.; Neville, P.R.; Barrett, H.A. Sinkhole Detection and Mapping Using Airborne LiDAR—A Practical Workflow. In *Tran-SET 2020*; American Society of Civil Engineers: Reston, VA, USA, 2020; pp. 100–111.
27. Miao, X.; Qiu, X.; Wu, S.S.; Luo, J.; Gouzie, D.R.; Xie, H. Developing efficient procedures for automated sinkhole extraction from Lidar DEMs. *Photogramm. Eng. Remote Sens.* **2013**, *79*, 545–554. [CrossRef]
28. Hofierka, J.; Gallay, M.; Bandura, P.; Šašak, J. Identification of karst sinkholes in a forested karst landscape using airborne laser scanning data and water flow analysis. *Geomorphology* **2018**, *308*, 265–277. [CrossRef]
29. Kim, Y.J.; Nam, B.H.; Youn, H. Sinkhole detection and characterization using LiDAR-derived DEM with logistic regression. *Remote Sens.* **2019**, *11*, 1592. [CrossRef]
30. Petschko, H.; Zehner, M.; Fischer, P.; Goetz, J. Terrestrial and Airborne Structure from Motion Photogrammetry Applied for Change Detection within a Sinkhole in Thuringia, Germany. *Remote Sens.* **2022**, *14*, 3058. [CrossRef]
31. Panno, S.V.; Luman, D.E. Characterization of cover-collapse sinkhole morphology on a groundwater basin-wide scale using lidar elevation data: A new conceptual model for sinkhole evolution. *Geomorphology* **2018**, *318*, 1–17. [CrossRef]
32. Rahimi, M.; Alexander, C. Locating sinkholes in LiDAR coverage of a glacio-fluvial karst, Winona County, MN. In *Full Proceedings of the Thirteenth Multidisciplinary Conference on Sinkholes and the Engineering and Environmental Impacts of Karst*; National Cave and Karst Research Institute: Carlsbad, NM, USA, 2013.
33. Wu, Q.; Deng, C.; Chen, Z. Automated delineation of karst sinkholes from LiDAR-derived digital elevation models. *Geomorphology* **2016**, *266*, 1–10. [CrossRef]
34. Zhu, J.; Taylor, T.; Currens, J.; Crawford, M. Improved karst sinkhole mapping in Kentucky using LiDAR techniques: A pilot study in Floyds Fork watershed. *J. Cave Karst Stud.* **2014**, *76*, 207–216. [CrossRef]
35. Wu, Q.; Liu, H.; Wang, S.; Yu, B.; Beck, R.; Hinkel, K. A localized contour tree method for deriving geometric and topological properties of complex surface depressions based on high-resolution topographical data. *Int. J. Geogr. Inf. Sci.* **2015**, *29*, 2041–2060. [CrossRef]
36. Zhu, J.; Pierskalla, W.P., Jr. Applying a weighted random forests method to extract karst sinkholes from LiDAR data. *J. Hydrol.* **2016**, *533*, 343–352. [CrossRef]

37. Jiang, Z.D.; Hu, S.; Deng, H.; Wang, N.L.; Zhang, F.Y.; Wang, L.; Wu, S.B.; Wang, X.G.; Cao, Z.W.; Chen, Y.X.; et al. Detection and automatic identification of loess sinkholes from the perspective of LiDAR point clouds and deep learning algorithm. *Geomorphology* **2024**, *465*, 109404–109417. [CrossRef]
38. Li, X.A.; Wang, L.; Hong, B.; Li, L.C.; Liu, J.; Lei, H.N. Erosion characteristics of loess tunnels on the Loess Plateau: A field investigation and experimental study. *Earth Surf. Process. Landf.* **2020**, *45*, 1945–1958. [CrossRef]
39. Jordan, M.I.; Mitchell, T.M. Machine learning: Trends, perspectives, and prospects. *Science* **2015**, *349*, 255–260. [CrossRef] [PubMed]
40. Kanevski, M.; Parkin, R.; Pozdnukhov, A.; Timonin, V.; Maignan, M.; Demyanov, V.; Canu, S. Environmental data mining and modeling based on Machine Learning algorithms and geostatistics. *Environ. Model. Softw.* **2004**, *19*, 845–855. [CrossRef]
41. Taheri, K.; Shahabi, H.; Chapi, K.; Shirzadi, A.; Gutiérrez, F.; Khosravi, K. Sinkhole susceptibility mapping: A comparison between bayes-based machine learning algorithms. *Land Degrad. Dev.* **2019**, *30*, 730–745. [CrossRef]
42. Zhu, J.; Nolte, A.M.; Jacobs, N.; Ye, M. Using machine learning to identify karst sinkholes from LiDAR-derived topographic depressions in the Bluegrass Region of Kentucky. *J. Hydrol.* **2020**, *588*, 125049. [CrossRef]
43. Verachtert, E.; Van Den Eeckhaut, M.; Poesen, J.; Govers, G.; Deckers, J. Prediction of spatial patterns ofcollapsed pipes in loess-derived soils in a temperate humid climate using logistic regression. *Geomorphology* **2011**, *130*, 185–196. [CrossRef]
44. Hosseinalizadeh, M.; Kariminejad, N.; Rahmati, O.; Keesstra, S.; Alinejad, M. How can statistical and artificial intelligence approaches predict piping erosion susceptibility? *Sci. Total Environ.* **2019**, *646*, 1554–1566. [CrossRef]
45. Peyman, A.; Mohammad, A.G.; Ahmet, R.M.; Ali, D.; Mohammad, A.L. Prediction of sinkhole hazard using artificial intelligence model with soil characteristics and GPR data in arid alluvial land in Central Iran. *Environ. Earth Sci.* **2023**, *85*, 372.
46. Poesen, J. Conditions for gully formation in the Belgian loam belt and some ways to control them. *Soil Technol.* **1989**, *1*, 39–52.
47. Faulkner, H. Badlands inmarl lithologies: A field guide to soil dispersion, subsurface erosion and piping-origin gullies. *Catena* **2013**, *106*, 42–53. [CrossRef]
48. Wilson, G.V.; Rigby, J.R.; Dabney, S.M. Soil pipe collapses in a loess pasture of Goodwin Creek watershed, Mississippi: Role of soil properties and past land use. *Earth Surf. Process. Landf.* **2015**, *40*, 1448–1463. [CrossRef]
49. Nasiri Aghdam, I.; Varzandeh, M.H.M.; Pradhan, B. Landslide susceptibility mapping using an ensemble statistical index (Wi) and adaptive neuro-fuzzy inference system (ANFIS) model at Alborz Mountains (Iran). *Environ. Earth Sci.* **2016**, *75*, 553. [CrossRef]
50. Geng, H.P.; Xu, W.Y.; Zheng, W.S.; Gao, X.L.; Pan, B.T. A hybrid mechanism for the initiation and expansion of loess caves across the Chinese Loess Plateau. *Land Degrad. Dev.* **2023**, *34*, 3329–3339. [CrossRef]
51. Intrieri, E.; Fontanelli, K.; Bardi, F.; Marini, F.; Fanti, R. Definition of sinkhole triggers and susceptibility based on hydrogeomorphological analyses. *Environ. Earth Sci.* **2018**, *77*, 4. [CrossRef]
52. Wood, N.J.; Doctor, D.H.; Alder, J.; Jones, J. Current and future sinkhole susceptibility in karst and pseudokarst areas of the conterminous United States. *Front. Earth Sci.* **2023**, *11*, 1207689. [CrossRef]
53. Baker, D. Quantitative parsimony and explanatory power. *Br. J. Philos. Sci.* **2003**, *54*, 245–259. [CrossRef]
54. Shatz, I. Why You Should Prefer Simpler Explanations, Parsimony. 25 June 2019. Available online: https://effectiviology.com/parsimony (accessed on 12 April 2024).
55. Subedi, P.; Subedi, K.; Thapa, B.; Subedi, P. Sinkhole susceptibility mapping in Marion County, Florida: Evaluation and comparison between analytical hierarchy process and logistic regression based approaches. *Sci. Rep.* **2019**, *9*, 7140. [CrossRef]
56. Nam, B.H.; Kim, Y.J.; Youn, H. Identification and quantitative analysis of sinkhole contributing factors in Florida's Karst. *Eng. Geol.* **2020**, *271*, 105610. [CrossRef]
57. Hitouri, S.; Meriame, M.; Ajim, A.S.; Pacheco, Q.R.; Nguyen-Huy, T.; Bao, P.Q.; ElKhrachy, I.; Varasano, A. Gully erosion mapping susceptibility in a Mediterranean environment: A hybrid decision-making model. *Int. Soil Water Conserv. Res.* **2024**, *12*, 279–297. [CrossRef]
58. Florinsky, I.V. Computation of the third-order partial derivatives from a digital elevation model. *Int. J. Geogr. Inf. Sci.* **2009**, *23*, 213–231. [CrossRef]
59. Kinnell, P.I.A. Alternative Approaches for Determining the USLE-M Slope Length Factor for Grid Cells. *Soil Sci. Soc. Am. J.* **2005**, *69*, 674–680. [CrossRef]
60. Boehner, J.; Selige, T. Spatial Prediction of Soil Attributes Using Terrain Analysis and Climate Regionalisation. In *SAGA—Analysis and Modelling Applications*; Boehner, J., McCloy, K.R., Strobl, J., Eds.; Goettinger Geographische Abhandlungen: Göttingen, Germany, 2006; Volume 115, pp. 13–27.
61. Passalacqua, P.; Do Trung, T.; Foufoula-Georgiou, E.; Sapiro, G.; Dietrich, W.E. A geometric framework for channel network extraction from lidar: Nonlinear diffusion and geodesic paths. *J. Geophys. Res. Earth Surf.* **2010**, *115*, 1–18. [CrossRef]
62. Quinn, P.F.; Beven, K.J.; Chevallier, P.; Planchon, O. The prediction of hillslope flow paths for distributed hydrological modelling using digital terrain models. *Hydrol. Process.* **1991**, *5*, 59–79. [CrossRef]
63. Gruber, S.; Peckham, S. Land-Surface Parameters and Objects in Hydrology. *Dev. Soil Sci.* **2008**, *33*, 293–308.
64. Skentos, A. Topographic position index based landform analysis of Messaria (Ikaria Island, Greece). *Acta Geobalcanica* **2018**, *4*, 7–15. [CrossRef]
65. Kołodyńska-Gawrysiak, R.; Gawrysiak, L.; Poesen, J.; Plak, A. The impact of natural closed depressions on soil organic carbon storage in eroded loess landscapes of East Poland. *Land Degrad. Dev.* **2023**, *34*, 5273–5292. [CrossRef]

66. Krawchuk, M.A.; Haire, S.L.; Coop, J.; Parisien, M.A.; Whitman, E.; Chong, G.; Miller, C. Topographic and fire weather controls of fire refugia in forested ecosystems of northwestern North America. *Ecosphere* **2016**, *7*, e01632. [CrossRef]
67. Riley, S.J.; De Gloria, S.D.; Elliot, R. A Terrain Ruggedness that Quantifies Topographic Heterogeneity. *Intermt. J. Sci.* **1999**, *5*, 23–27.
68. Heerdegen, R.G.; Beran, M.A. Quantifying source areas through land surface curvature. *J. Hydrol.* **1982**, *57*, 359–373. [CrossRef]
69. Jasiewicz, J.; Stepinski, T.F. Geomorphons—A pattern recognition approach to classification and mapping of landforms. *Geomorphology* **2013**, *182*, 147–156. [CrossRef]
70. Wood, J. Geomorphometry in LandSerf. *Dev. Soil Sci.* **2009**, *33*, 333–349.
71. Ballabio, C.; Sterlacchini, S. Support vector machines for landslide susceptibility mapping: The staffora River Basin case study. *Italy Math. Geosci.* **2012**, *44*, 47–70. [CrossRef]
72. Tien Bui, D.; Pradhan, B.; Lofman, O.; Revhaug, I. Landslide susceptibility as-sessment in vietnam using support vector machines, decision tree, and naïve bayes models. *Math. Probl. Eng.* **2012**, *2012*, 1–26. [CrossRef]
73. Vapnik, V. *Statistical Learning Theory*; John Wiley and Sons, Inc.: New York, NY, USA, 1998.
74. Kavzoglu, T.; Colkesen, I. A kernel functions analysis for support vector machines for land cover classification. *Int. J. Appl. Earth Obs. Geoinf.* **2009**, *11*, 352–359. [CrossRef]
75. Abe, S. *Support Vector Machines for Pattern Classification*, 2nd ed.; Springer: London, UK, 2010.
76. Smola, A.J.; Schölkopf, B. A tutorial on support vector regression. *Stat. Comput.* **2004**, *14*, 199–222. [CrossRef]
77. Tien Bui, D.; Anh Tuan, T.; 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]
78. Al-Juaidi, A.E.M.; Nassar, A.M.; Al-Juaidi, O.E.M. Evaluation of flood susceptibility mapping using logistic regression and GIS conditioning factors. *Arab. J. Geosci.* **2018**, *11*, 765. [CrossRef]
79. Ozdemir, A. Sinkhole susceptibility mapping using logistic regression in Karapınar (Konya, Turkey). *Bull. Eng. Geol. Environ.* **2016**, *75*, 681–707. [CrossRef]
80. Ciotoli, G.; Di Loreto, E.; Finoia, M.G.; Liperi, L.; Meloni, F.; Nisio, S.; Sericola, A. Sinkhole susceptibility, Lazio Region, Central Italy. *J. Maps* **2016**, *12*, 287–294. [CrossRef]
81. Krizhevsky ASutskever, I.; Hinton, G.E. ImagNet classification with deep convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems (NIPS), Lake Tahoe, NV, USA, 3–6 December 2012; pp. 1106–1114.
82. Lecun, Y.; Bottou, L.; Gengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
83. Tien Bui, D.; Nguyen, Q.P.; Hoang, N.D.; Klempe, H. A novel fuzzy K-nearest neighbor inference model with differential evolution for spatial prediction of rainfall-induced shallow landslides in a tropical hilly area using GIS. *Landslides* **2017**, *14*, 1–17. [CrossRef]
84. Zhang, Y.; Cao, G.; Wang, B.; Li, X. A novel ensemble method for k-nearest neighbor. *Pattern Recogn.* **2019**, *85*, 13–25. [CrossRef]
85. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
86. Youssef, A.M.; Pourghasemi, H.R.; Pourtaghi, Z.S.; Al-Katheeri, M.M. Landslide susceptibility mapping using random forest, boosted regression tree, classification and regression tree, and general linear models and comparison of their performance at Wadi Tayyah Basin, Asir Region Saudi Arabia. *Landslides* **2016**, *13*, 839–856. [CrossRef]
87. Chen, W.; Shirzadi, A.; Shahabi, H.; Ahmad, B.B.; Zhang, S.; Hong, H.; Zhang, N. A novel hybrid artificial intelligence approach based on the rotation forest ensemble and naïve Bayes tree classifiers for a landslide susceptibility assessment in Langao County, China. *Geomat. Nat. Hazards Risk* **2017**, *8*, 1955–1977. [CrossRef]
88. Lagomarsino, D.; Tofani, V.; Segoni, S.; Catani, F.; Casagli, N. A tool for classifi-cation and regression using random forest methodology: Applications to landslide susceptibility mapping and soil thickness modeling. *Environ. Model. Assess.* **2017**, *22*, 201–214. [CrossRef]
89. Hong, H.; Liu, J.; Tien Bui, D.; Pradhan, B.; Acharya, T.D.; Pham, B.T.; Zhu, A.-X.; Chen, W.; Ahmad, B.B. Landslide susceptibility mapping using J48 Decision Tree with AdaBoost, Bagging and Rotation Forest ensembles in the Guangchang area (China). *CATENA* **2018**, *163*, 399–413. [CrossRef]
90. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
91. Chung CJ, F.; Fabbri, A.G. Validation of spatial prediction modelsfor landslidehazard mapping. *Nat. Hazards* **2003**, *30*, 451–472. [CrossRef]
92. Lee, E.J.; Shin, S.Y.; Ko, B.C.; Chang, C. Early sinkhole detection using a drone-based thermal camera and image processing. *Infrared Phys. Technol.* **2016**, *78*, 223–232. [CrossRef]
93. Silva, O.L.; Bezerra, F.H.; Maia, R.P.; Cazarin, C.L. Karst landforms revealed at various scales using LiDAR and UAV in semi-arid Brazil: Consideration on karstification processes and methodological constraints. *Geomorphology* **2017**, *295*, 611–630. [CrossRef]
94. Suh, J.; Choi, Y. Mapping hazardous mining-induced sinkhole subsidence using unmanned aerial vehicle (drone) photogrammetry. *Environ. Earth Sci.* **2017**, *76*, 144. [CrossRef]
95. Gutiérrez, F.; Galve, J.P.; Lucha, P.; Castañeda, C.; Bonachea, J.; Guerrero, J. Integrating geomorphological mapping, trenching, InSAR and GPR for the identification and characterization of sinkholes: A review and application in the mantled evaporite karst of the Ebro Valley (NE Spain). *Geomorphology* **2011**, *134*, 144–156. [CrossRef]

96. Liu, R.; Sun, H.; Qin, J.; Zheng, Z. A multi-geophysical approach to assess potential sinkholes in an urban area. *Eng. Geol.* **2023**, *318*, 107100. [CrossRef]
97. Guo, Z.T.; Ding, Z.L.; Liu, T.S. The deposition in loess-soil forming and the quaternary weather. *Chin. Sci. Bull.* **1996**, *41*, 56–59. (In Chinese)
98. Xu, L.; Dai, F.C. Statistical analysis of the characteristic parameters of loess landslides at the South Jingyang Plateau. *Hydrogeol. Eng. Geol.* **2008**, *5*, 28–32.
99. Wang, G.H.; Zhang, D.X.; Furuya, G.; Yang, J. Pore-pressure generation and fluidization in a loess landslide triggeredby the 1920 Haiyuan earthquake, China: A case study. *Eng. Geol.* **2014**, *174*, 36–45. [CrossRef]
100. Gutiérrez, F.; Guerrero, J.; Lucha, P. Quantitative sinkhole hazard assessment. A case study from the Ebro Valley evaporite alluvial karst (NE Spain). *Nat. Hazards* **2008**, *45*, 211–233. [CrossRef]
101. Taheri, K.; Gutierrez, F.; Mohseni, H.; Raeisi, E.; Taheri, M. Sinkhole susceptibility mapping using the analytical hierarchy process (AHP) and magnitude–frequency relationships: A case study in Hamadan 1170 province, Iran. *Geomorphology* **2015**, *234*, 64–79. [CrossRef]
102. Qiu, X.; Wu, S.-S.; Chen, Y. Sinkhole susceptibility assessment based on morphological, imagery, and contextual attributes derived from GIS and imagery data. *J. Cave Karst Stud.* **2020**, *82*, 1125. [CrossRef]
103. Ozdemir, A. Sinkhole susceptibility mapping using a frequency ratio method and GIS technology near Karapınar, Konya-Turkey. *Procedia Earth Planet. Sci.* **2015**, *15*, 502–506. [CrossRef]
104. Kim, Y.J.; Nam, B.H.; Jung, Y.H.; Liu, X.; Choi, S.; Kim, D.; Kim, S. Probabilistic spatial susceptibility modeling of carbonate karst sinkhole. *Eng. Geol.* **2022**, *306*, 106728. [CrossRef]
105. Nachappa, T.G.; Piralilou, S.T.; Gholamnia, K.; Ghorbanzadeh, O.; Rahmati, O.; Blaschke, T. Flood susceptibility mapping with machine learning, multi-criteria decision analysis and ensemble using Dempster Shafer Theory. *J. Hydrol.* **2020**, *590*, 125275. [CrossRef]
106. Tehrani, F.S.; Calvello, M.; Liu, Z.; Zhang, L.; Lacasse, S. Machine learning and landslide studies: Recent advances and applications. *Nat. Hazards* **2022**, *114*, 1197–1245. [CrossRef]
107. Chen, Y.Z.; Chen, W.; Janizadeh, S.; Bhunia, G.S.; Bera, A.; Pham, Q.B.; Linh NT, T.; Balogun, A.L.; Wang, X.J. Deep learning and boosting framework for piping erosion susceptibility modeling: Spatial evaluation of agricultural areas in the semi-arid region. *Geocarto Int.* **2022**, *37*, 4628–4654. [CrossRef]
108. Basso, A.; Bruno, E.; Parise, M.; Pepe, M. Morphometric analysis of sinkholes in a karst coastal area of southern Apulia (Italy). *Environ. Earth Sci.* **2013**, *70*, 2545–2559. [CrossRef]
109. Jouves, J.; Viseur, S.; Arfib, B.; Baudement, C.; Camus, H.; Collon, P.; Guglielmi, Y. Speleogenesis, geometry, and topology of caves: A quantitative study of 3D karst conduits. *Geomorphology* **2017**, *298*, 86–106. [CrossRef]
110. Li, X.A.; Song, Y.X.; Ye, W.J. *Engineering Geological Research on Tunnel-Erosion in Loess*; Tongji University Press: Shanghai, China, 2010. (In Chinese)

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

Flooded Infrastructure Change Detection in Deeply Supervised Networks Based on Multi-Attention-Constrained Multi-Scale Feature Fusion

Gang Qin [1,2,3], Shixin Wang [1,2,3], Futao Wang [1,2,3,*], Suju Li [2,4], Zhenqing Wang [1,2,3], Jinfeng Zhu [1,2], Ming Liu [2,4], Changjun Gu [2,4] and Qing Zhao [1,2]

[1] Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China; qingang20@mails.ucas.ac.cn (G.Q.)
[2] Key Laboratory of Emergency Satellite Engineering and Application, Ministry of Emergency Management, Beijing 100124, China
[3] University of Chinese Academy of Sciences, Beijing 100094, China
[4] National Disaster Reduction Center of China, No. 6 Guangbaidong Road, Chaoyang District, Beijing 100124, China
* Correspondence: wangft@aircas.ac.cn

Abstract: Flood disasters are frequent, sudden, and have significant chain effects, seriously damaging infrastructure. Remote sensing images provide a means for timely flood emergency monitoring. When floods occur, emergency management agencies need to respond quickly and assess the damage. However, manual evaluation takes a significant amount of time; in current, commercial applications, the post-disaster flood vector range is used to directly overlay land cover data. On the one hand, land cover data are not updated in time, resulting in the misjudgment of disaster losses; on the other hand, since buildings block floods, the above methods cannot detect flooded buildings. Automated change-detection methods can effectively alleviate the above problems. However, the ability of change-detection structures and deep learning models for flooding to characterize flooded buildings and roads is unclear. This study specifically evaluated the performance of different change-detection structures and different deep learning models for the change detection of flooded buildings and roads in very-high-resolution remote sensing images. At the same time, a plug-and-play, multi-attention-constrained, deeply supervised high-dimensional and low-dimensional multi-scale feature fusion (MSFF) module is proposed. The MSFF module was extended to different deep learning models. Experimental results showed that the embedded MSFF performs better than the baseline model, demonstrating that MSFF can be used as a general multi-scale feature fusion component. After FloodedCDNet introduced MSFF, the detection accuracy of flooded buildings and roads changed after the data augmentation reached a maximum of 69.1% MIoU. This demonstrates its effectiveness and robustness in identifying change regions and categories from very-high-resolution remote sensing images.

Keywords: change detection; deep learning; multi-scale feature fusion; flooded buildings and roads; very-high-resolution remote sensing images

1. Introduction

Floods are one of the major natural disasters, often characterized by frequent, sudden occurrences and significant cascading effects driven by global climate change [1]. Floods cause huge loss of life; damage to infrastructure and property; and other environmental issues, including post-disaster cleanup and reconstruction, and healthcare [2]. To mitigate the immediate impacts of flooding, first responders need to respond quickly and assess the damage during a flood, or shortly after a flood occurs [3,4]. The most important of these needs is that of assessing the damage to buildings and roads to determine where help is

needed, and which roads are blocked by flooding [5]. Remote sensing images are the main tool for quickly mapping flood-prone areas [6,7]. However, detailed emergency monitoring of the disaster situation and damage assessment take a great deal of time. During flood disasters, the inundation of building roofs by floodwaters results in conventional flood and building detection methods not being able to inspect flooded buildings. The complex characteristics of the fully submerged road also influence the results of traditional road detection methods, causing them to decline. Therefore, it is of great significance to use deep learning models to quickly and accurately quantify damaged buildings and roads.

At present, in order to determine flooded buildings and flooded roads in operational applications for disaster responses by the government, the solution is to use post-disaster flood vector maps to directly overlay land cover data. On the one hand, there is often a time lag in the acquisition of effective optical imagery in disaster areas. At this time, roofs are not inundated by partially receding floodwaters, and the flooded roads show a complex state. On the other hand, untimely updated land cover data can lead to the misclassification of disaster losses. Information is obtained through flood-, building-, and road-detection methods. As a consequence of the above, a reliance on three types of information to obtain results for flooded buildings and flooded roads can lead to a large number of omissions. Dual-temporal image change-detection technology can accurately determine the damage to infrastructure before and after floods. By analyzing and comparing geographic information, remote sensing images and other data from different periods, remote sensing image change detection plays an important role in geographic data updating, disaster damage assessment, and other aspects. Compared with change detection for normal-state buildings and roads, the change detection of damaged infrastructure is more challenging. There are additional interfering features in disaster scenarios such as damaged non-building structures in earthquake disasters [8,9], and the mixtures of flood water from various landforms in flood disasters [10,11]. The detection of damaged infrastructure in earthquake disasters has been widely studied. However, change detection for flooded buildings and roads is a difficult task. Unlike earthquakes, that cause complete destruction and severe collapse, most of the flooded buildings and roads are still in a normal condition when viewed from the air during floods. Therefore, there is an urgent need for feature extractors with stronger context-awareness and change-detection structures that can fully mine the dual-temporal semantic information before and after floods. With the introduction of very-high-resolution flooded buildings and roads datasets, such as FloodNet [12], SpaceNet8 [5], and xBD [13], change-detection structures have been initially applied to the detection of flooded buildings and roads in flood scenarios [14]. However, different change-detection structures have different abilities in characterizing changes in flooded buildings and roads. The different change-detection structures are shown in Figure 1, and include, as follows: early fusion (a); middle fusion (b); late fusion (c); and first segmentation and then change detection (d). It is still unclear which change-detection structure has the strongest ability.

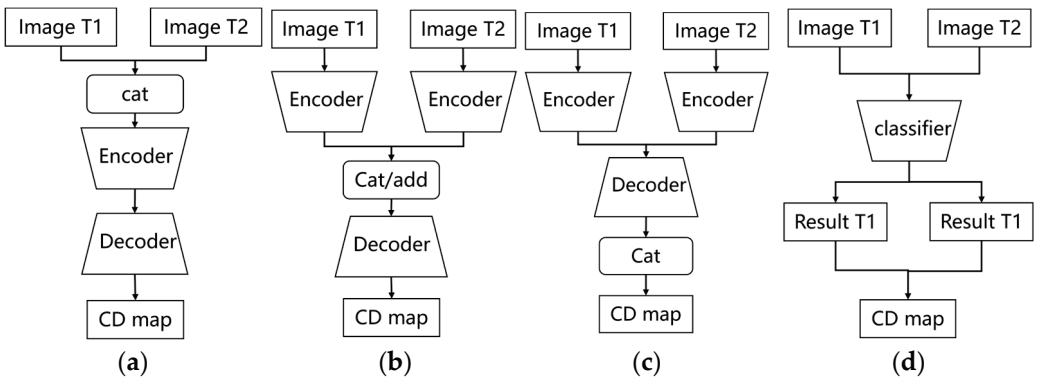


Figure 1. Common change-detection structures. From left to right: (**a**–**d**). (**a**) early fusion, (**b**) middle fusion, (**c**) late fusion, (**d**) first segmentation and then change detection.

Deep learning models are very good at discovering high-dimensional structures in complex data [15]. The models use a convolutional structure to capture spatial context information and play an important role in disaster emergency monitoring and damage assessment. Refs. [16,17] pioneered the application of convolutional neural networks to image semantic segmentation and innovated on the encoding–decoding structure to improve the upsampling effect by preserving the pooling index. In subsequent studies, such as those conducted by Refs. [18–21], multi-scale contextual information was effectively integrated through U-shaped structures, pyramid pooling modules, dilated convolutions, and multi-branch parallel convolutions, etc., enhancing the model's ability to understand complex scenes. However, the above studies still have some limitations with regard to reducing complexity and generating multi-scale features with more powerful capabilities, while fully considering the use of scale-related contextual information. Refs. [22–24] proposed an NAS search strategy to build the network structure and a residual structure to deepen the network and transformerize the self-attention mechanism of the convolutional neural network to enhance the feature extraction capability of deep learning models while preventing their complexity from exceeding the memory limit. Encoders built relying on these design ideas significantly enhance the change-detection capabilities of deep learning models.

The semantic features extracted from the feature extractor, such as the low-dimensional features, focus on detailed information such as edges and textures; high-dimensional features focus on more abstract semantic information. Multi-scale feature fusion and layer feature fusion strategies are designed to realize the fusion of high-dimensional features and low-dimensional features; thus, a large number of channels are introduced into the fusion strategy [25,26]. The multi-channel convolution kernel has significant spatial similarity and channel redundancy [27]. To alleviate channel redundancy, SE and ECA focus on the attention of the channel dimension and assign appropriate weights to feature channels at different levels [28]. Since convolutional kernels produce similar responses to similar patterns in the local space, in studies by Refs. [29–31], CBAM and scSE simultaneously consider attention in the spatial and channel dimensions to further solve the spatial independence problem. Some studies embed attention mechanisms in the encoder or decoder of the network structure [32–34], ignoring the fact that multi-scale features in high-dimensional and low-dimensional spaces are independent of each other, and the utilization of multi-scale features in change-detection tasks is relatively limited. Dynamic multi-scale feature fusion or multi-attention constraint methods are designed to learn fusion weights from features and guide the organic fusion of shallow and deep semantic features [35,36]. These multi-scale feature fusions, based on the attention mechanism, depend on specific deep learning models—the above modules cannot be directly applied in deep learning models with different structures.

In summary, in flood emergency monitoring and disaster assessment tasks, professional researchers in the current operational application field use flood vector maps to overlay land cover data. However, due to the fact that land cover data are not updated in time or because the buildings cover the floods, the monitoring accuracy of changes in flooded buildings and roads is not high. This study designs a change-detection structure, FloodedCD, that uses dual-temporal images before and after floods. Different deep learning models are tested for their change-detection ability in flooded buildings and roads in very-high-resolution remote sensing images. As deep learning models have different modeling capabilities in complex flood scenarios, this study confirms the superiority of FloodedCDNet for road change detection in flooded buildings. Facing the problems of channel redundancy and spatial similarity in the fusion of high-dimensional and low-dimensional feature maps of different sizes, this study constructed a plug-and-play multi-attention-constrained multi-scale feature extraction module and used a deep supervised learning strategy to make full use of the global–local multi-scale features of flood scenes to improve the accuracy of change detection. The FloodedCDNet proposed in this study improves

the ability to monitor changes in flooded buildings and roads in complex flood scenarios, confirming its excellence in flood scenarios.

2. Methods

This study proposes the model (FloodedCDNet) workflow shown in Figure 2, which acquires the change detection of flooded buildings and roads from ultra-high-resolution remote sensing imagery. It learns five categories of change information through the encoder–decoder architecture. By extracting the changing features of normal buildings and roads in pre-disaster images, and flooded buildings and roads in post-disaster images, this change-detection structure improves the identification of flooded buildings and roads. The feature extractor constructed in this study adopts a Siamese network based on ConvNeXt. With strong contextual awareness, the semantic features of flooded buildings and roads are obtained and the characteristics of buildings, roads, and floods from dual-temporal images can be extracted. In order to strengthen the extraction and fusion of semantic features at different levels of dual-temporal images, this study constructs a deeply supervised, high-dimensional–low-dimensional multi-scale feature fusion module with multi-attention constraints that can be used immediately. Robust semantic representations of the global–local infrastructure and flooding are obtained, and the accuracy of the rapid extraction of flooded buildings and roads is significantly improved. Subsequently, the data augmentation strategy and the deep supervision module for supervising the hidden layer are gradually introduced to improve detection accuracy.

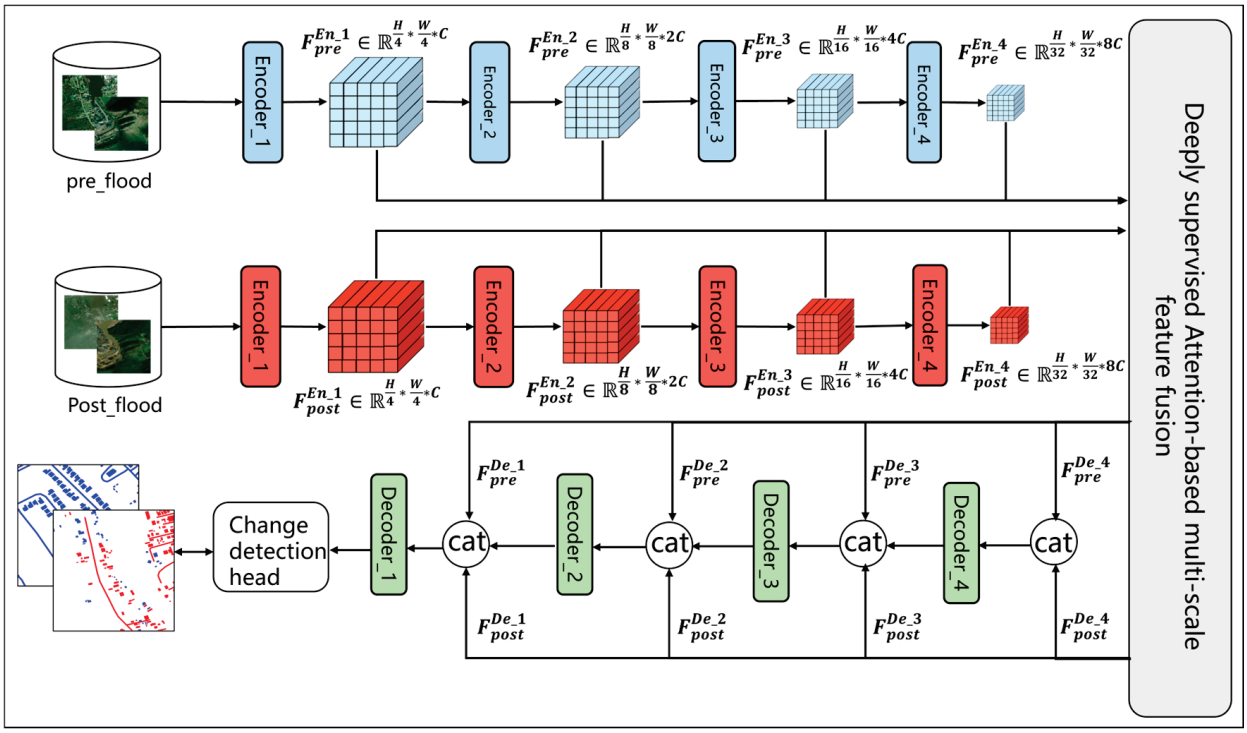


Figure 2. The overall network structure of FloodedCDNet.

FloodedCDNet uses a feature fusion change-detection structure with shared weights. When the pre-disaster image is used as input, the encoder outputs the feature map of $F_{pre}^{En\_1}$, $F_{pre}^{En\_2}$, $F_{pre}^{En\_3}$ and $F_{pre}^{En\_4}$ at each stage. When the post-disaster image is used as input, the encoder outputs the feature map of $F_{post}^{En\_1}$, $F_{post}^{En\_2}$, $F_{post}^{En\_3}$ and $F_{post}^{En\_4}$ at each stage. The above feature maps are fused by the MSFF module to obtain $F_{pre}^{De\_1}$, $F_{pre}^{De\_2}$, $F_{pre}^{De\_3}$, $F_{pre}^{De\_4}$, $F_{post}^{De\_1}$, $F_{post}^{De\_2}$, $F_{post}^{De\_3}$ and $F_{post}^{De\_4}$. Finally, after the weight calculation of the change-detection head, the model outputs the information of the buildings and roads damaged by the flood.

2.1. Detailed Explanation of Change-Detection Structure

Deep learning is widely used in change detection due to its excellent feature extraction capabilities. There are many different change-detection structures based on deep learning. DL-based change-detection frameworks can be divided into single-branch and dual-branch structures, according to their feature extraction strategies. The single-branch structure uses early fusion to merge temporal images into multi-spectral images. However, due to the lack of independent original input images, it is difficult for the single-branch structure to extract the fine temporal features of each dual-temporal image, which is not conducive to the subsequent change detection of the network. The dual-branch structure is a Siamese network consisting of two sub-networks with shared weights [37,38] or non-shared weights [39,40], and has two dual-phase fusion modes, mid-term fusion and late-term fusion. Each branch independently learns high-level feature representations of single-phase images through multiple convolutions. This powerful representation ability enables the dual-branch network to effectively capture the change information in complex RS images, thereby improving the accuracy of change detection. Ref. [16] proposed two Siamese networks, namely, FC-Siam-conc and FC-Siam-diff, which adopt fully convolutional shared weights. In order to provide the model with stronger context perception ability for single temporal images, this study adopts the mid-term fusion strategy, embeds the MSFF module, and improves the mid-term fusion strategy. Based on [41–43], the change-detection structure FloodedCD was designed. It can extract the characteristics of buildings, roads, and floods from dual-temporal images. At the same time, it obtains the location information of flooded buildings and roads with strong contextual awareness. Figure 3 shows the simplified change-detection structure, encoder–decoder, and fusion method.

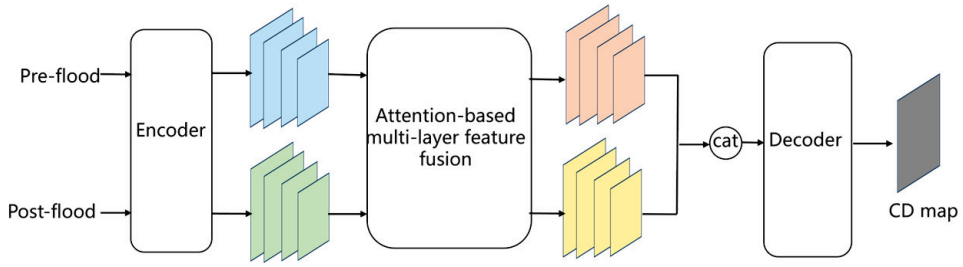


Figure 3. FloodedCD simplified change-detection structure diagram.

2.2. Detailed Description of Deep Learning Model

In order to balance the number of model parameters and the performance of feature extraction, this study introduced ConvNeXt-Tiny [24] as the backbone of the encoder and added ImageNet [44] pre-trained weights on it. The encoder, decoder, and high-dimensional and low-dimensional multi-scale feature fusion module were constructed as shown in Figure 4. FloodedCDNet's encoder consists of four stages. The input image is run through the encoder in four stages, in turn. Each stage consists of a different number of encoder modules, and the number of encoder modules is (3, 3, 9, 3). In order to adapt to the task of the change detection of dual-temporal images before and after the flood disaster, the encoder of FloodedCDNet is a twin type. The weights of the image-sharing encoder before and during the disaster are realized. The encoder structure is as follows: deep convolution; layernorm; point-by-point convolution; GELU; and point-by-point convolution [45–47]. The above encoder structure has a stronger context-awareness and can extract more accurate flood damage semantic information. FloodedCDNet's decoder also consists of four stages. The output feature maps of the previous stage decoder and the encoder of the four stages are fused and are then input, in turn, in the four-stage decoders. The decoder structure consists of convolution, inverse convolution, BN and ReLU, which realize the spatial and semantic feature fusion and complementation of high-dimensional and low-dimensional multi-scale features.

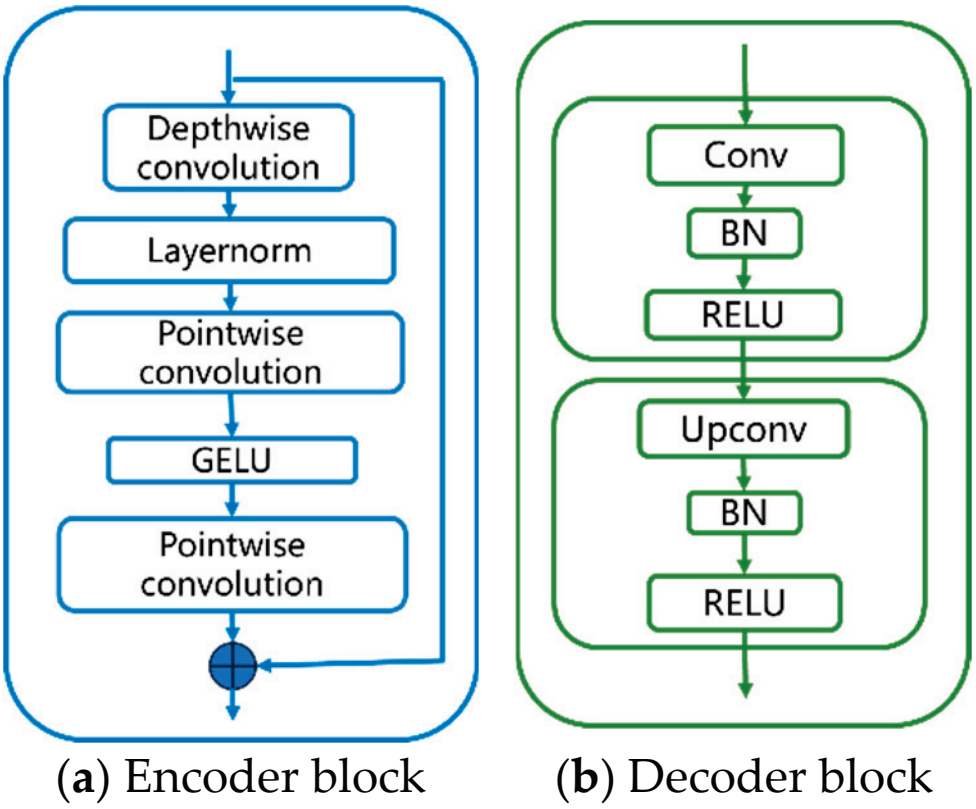


Figure 4. FloodedCDNet detailed network structure diagram. (**a**) is the specific details of the encoder module. (**b**) is the specific details of the decoder module.

2.3. Multi-Attention-Constrained High-Dimensional and Low-Dimensional Multi-Scale Feature Fusion

PSPNet first proposed pyramid-structured feature fusion [19]. In subsequent research work, feature fusion was widely used in image segmentation [48,49]. Even though Ref. [50] incorporates all fine-grained spatial information and coarse-grained semantic information into the decoder through full-scale skip connections, it cannot be used as a plug-and-play into various promising deep learning models and is heavily dependent on the specific structure of the module. ScSE is a joint-attention mechanism. It combines the channel attention mechanism and the spatial attention mechanism of the convolution module [51]. In order to overcome the problem of multi-scale feature fusion module migration and make full use of multi-scale features, this study constructed a plug-and-play, multi-attention-constrained, high-dimensional and low-dimensional multi-scale feature fusion module, which will significantly improve the performance of flooded building and road change-detection accuracy. As shown in Figure 5, $MSFF_{single}$ is based on the low-dimensional features $F_{pre}^{En\_1}$ and integrates the $F_{pre}^{En\_2}$, $F_{pre}^{En\_3}$ and $F_{pre}^{En\_4}$ features of other scales. First, $F_{pre}^{En\_2}$, $F_{pre}^{En\_3}$ and $F_{pre}^{En\_4}$ are upsampled to the same size as $F_{pre}^{En\_1}$. Then, the four scale feature maps are used along the channel dimension cat. The effective features are enhanced and the useless features are weakened by the channel spatial attention module. Finally the $F_{pre}^{De\_1}$ is obtained. $MSFF_{all}$ takes $F_{pre}^{En\_2}$, $F_{pre}^{En\_3}$, $F_{pre}^{En\_4}$, $F_{post}^{En\_1}$, $F_{post}^{En\_2}$, $F_{post}^{En\_3}$, and $F_{post}^{En\_4}$ as inputs, and follows the same fusion idea to obtain $F_{pre}^{De\_2}$, $F_{pre}^{De\_3}$, $F_{pre}^{De\_4}$, $F_{post}^{De\_1}$, $F_{post}^{De\_2}$, $F_{post}^{De\_3}$ and $F_{post}^{De\_4}$.

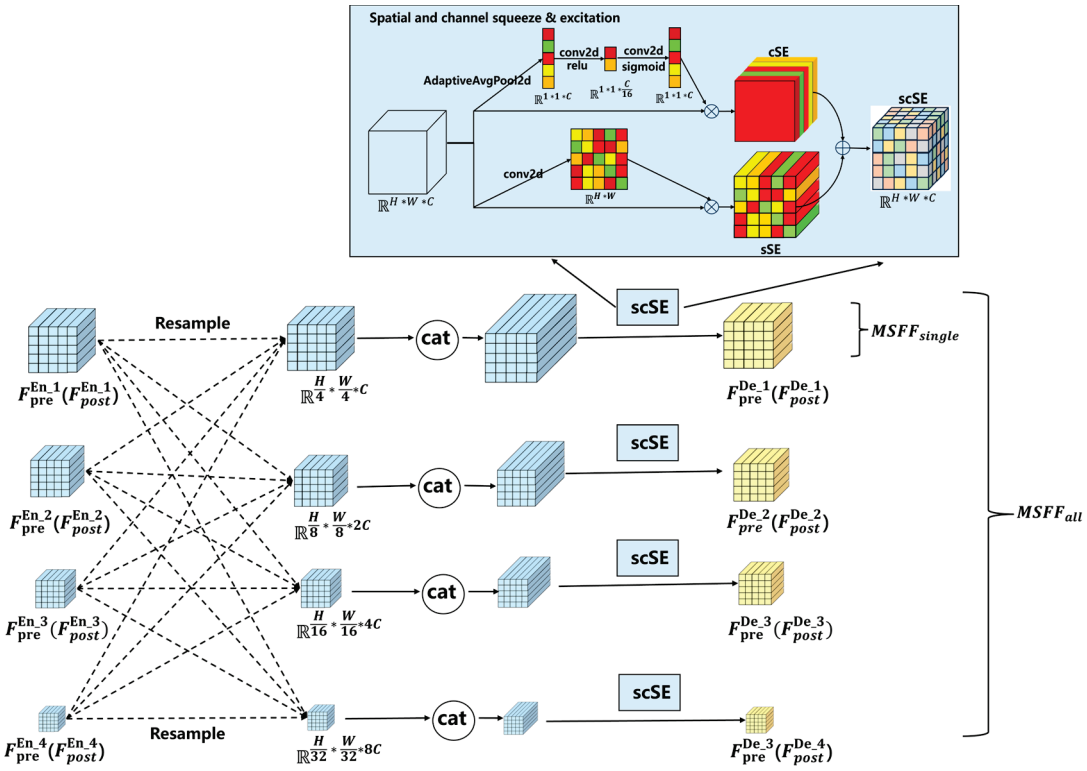


Figure 5. Multi-attention-constrained, high-dimensional and low-dimensional multi-scale feature fusion module diagram.

2.4. Deeply Supervised Learning

Different to cross-entropy loss [52], dice loss [53] and focal loss [54], which are used as the main loss functions of the network structure, Ref. [50] proposed the use of cross-entropy loss as the main loss function and focal loss as an auxiliary loss function to optimize the hidden layer feature extraction capability. This study draws on Ref. [55] by adding deep supervised learning to the hidden layers of the network and directly optimizes the gradient backpropagation process of the hidden layer weights to make them biased towards high-discriminative feature maps. They are useful to enhance the flood inundation feature extraction ability of the hidden layer in the network, thereby enhancing the flood and submergence feature extraction capability of the hidden layer in the network. Pre-flood images, post-flood images, and the water body masks are characterized as $D = \left\{\left(x_i^{pre}, x_i^{post}, p_i\right)\right\}$. In this study, the second decoder of FloodedCDNet is called De_2, and a deep supervised classifier $\mathcal{L}_{ce-stage2}(q_{b-De\_2}, p_i)$ is added. This introduces deep supervision optimization for the network's encoder and the first two decoders, which can prompt the model to learn features at these different levels and improve the detection results of flooded building and road changes.

$q_b = p_m\left(y\middle|x_i^{pre}, x_i^{post}; \mathcal{M}\right)$ is the heat map generated by the inference of model M on the training set $\left\{\left(x_i^{pre}, x_i^{post}, p_i\right)\right\}$. In this study, q_{b-De_2} is the feature map generated by De_2. B is the batchsize of samples in the training set; each p_i is the onehot label of the labeled data, and upsamples the feature map, as follows:

$$\mathcal{L}_{ce-stage2}(q_{b-De\_2}, p_i) = -\frac{1}{B}\sum_{b=1}^{B}\left\{p_i * \text{upsample}(q_{b-De\_2}) + (1-p_i) * \left(1 - \text{upsample}(q_{b-De\_2})\right)\right\} \quad (1)$$

3. Experiments

3.1. Data

Remote sensing can quickly and extensively obtain land surface information. It is able to accurately identify the extent of a flood. At the same time, remote sensing images can clearly show the damage to buildings and road infrastructure in flood disasters. The change-detection structure can realize the detection of flood-damaged buildings and roads. The above change-detection tasks require registered images and ultra-high-resolution remote sensing images (higher than 1 m) before and after the flood. The selection of research datasets is crucial for flood monitoring using remote sensing techniques; therefore, the following datasets are used in this study.

Spacenet8 Dataset: This dataset involves two hurricane disasters that occurred in Bonn, Germany and Louisiana, United States, during 2021, as well as a third hurricane disaster site used for testing. A total of 12 images were used for data collection, including pre-disaster and post-disaster images. Each image contains R, G, and B channels, and the spatial resolution ranges from 0.3 m to 0.8 m. Professionals marked all buildings and roads based on pre-disaster images, and then classified the flooded buildings and roads based on post-disaster images. The AOIs of the three regions cover a total area of 850 km^2, including 32,000 buildings and 1300 km of roads, of which 13% and 15% were flooded, respectively. In order to meet the data requirements of deep learning models, this study preprocessed the dataset labels according to the script provided in [5] and resampled the images to make them consistent in size before and after the disaster. Since the official dataset does not provide test set labels, this study divides all labeled datasets into training sets and test sets in a 4:1 ratio.

The data augmentation strategy makes a series of changes and enrichments to the limited data to increase the diversity and number of data samples [56]. There are few remote sensing images of flood disaster scenes. In order to expand the number of images in flood scenes, this study uses data augmentation strategies such as random cropping, horizontal inversion, vertical inversion, and normalization. When the dataset is limited, this can improve the generalization ability of the model, alleviate the overfitting problem, and increase the robustness of the model to different changes and disturbances.

3.2. Experimental Setup and Evaluation Metrics

In this article, all experiments were performed using PyTorch1.12.1. NVIDIA GeForce RTX 3090, designed in California, USA, is used to accelerate model training, with a video memory of 24G, and train 100 epochs each time.

Due to GPU memory limitations, we pre-processed the size of SpaceNet8 to 1300×1300 to unify the resolution and size of pre- and post-flood imagery for better training metrics. The batch size was set to 2. The data enhancement strategy adopted random crop, horizontal flip, vertical flip, and normalization. The Adam optimizer optimized the model, with the initial learning rate set to 1×10^{-4} and the learning rate reduced by 50% every 40 epochs of training. The weight of deep supervised learning was 0.4. This study used ImageNet to initialize the weights of the convolutional layers in the model.

The evaluation index IoU (intersection over union) is one of the commonly used indicators to measure the performance of image segmentation algorithms. It evaluates the accuracy and consistency of the segmentation results based on the ratio of the intersection and union between the true and predicted values. In the IoU calculation, we focused on four key parameters: TP, FP, TN, and FN. TP represents the positive samples predicted by the model as positive. FP represents the negative samples predicted by the model as positive. TN represents the negative samples predicted by the model as negative. FN represents the positive samples predicted by the model as negative.

The calculation formula of IoU is as follows:

$$IoU = \frac{TP}{TP + FP + FN} \tag{2}$$

$$MIoU = \frac{1}{k}\sum_{i=0}^{k} IoU_i \tag{3}$$

$$Precison = \frac{TP}{TP + FP},\ Recall = \frac{TP}{TP + FN} \tag{4}$$

$$F1_{Score} = \frac{2 * (Precision * Recall)}{Precision + Recall} \tag{5}$$

Its value range is between 0 and 1. The closer the value is to 1, the higher the match between the segmentation result and the true value.

3.3. Comparison Methods

To verify the advantages of the FloodedCDNet proposed in this study, we compared FloodedCDNet with several state-of-the-art deep learning models, including, as follows: FC-EF; FC-Siam-Conc; FC-Siam-Diff; EfficientNet; ResNeSt; SNUNet; ChangeFormer; and ConvNeXt.

In FC-EF [16], the EF architecture merges the two images before passing them through the network, treating them as different color channels and directly catenating them together; in the decoding stage, the spatial information is restored with the help of skip connections.

FC-Siam-Conc [16] first processes two images separately through network branches with shared structure and parameters and then fuses the feature maps of the two branches with the corresponding layers of the decoder through skip connections.

FC-Siam-Diff [16] mainly focuses on computing the absolute difference between the two branches, merges the two branches after the encoder and establishes skip connections with the corresponding decoder layers.

EfficientNet [23] conducted a systematic study on model expansion and proposed a new scaling method that uses a simple and efficient compound coefficient to uniformly scale all dimensions of depth/width/resolution. It also uses neural architecture search technology to design a new baseline network, thereby improving performance.

ResNeSt [57] proposed a modular architecture that applied channel-wise attention on different network branches to leverage their success in capturing cross-feature interactions and learning diverse representations. Excellent transfer results were achieved on public benchmark datasets.

SNUNet [37] is a densely connected Siamese network for change detection. It realizes multi-scale feature transmission between the encoder and decoder and decoder and decoder and proposes an ensemble channel attention module (ECAM) for deep supervision.

The method proposed by ChangeFormer [58] unifies the hierarchical transformer encoder and the multi-layer perceptron (MLP) decoder in a Siamese network architecture to effectively render the multi-scale and long-range details required for accurate change detection.

ConvNeXt [24] tests the limits of what is possible with pure ConvNets. We designed the visual transformer by incrementally "transformerizing" the standard ResNet; in the process, we discovered several key components that made a difference to the performance. ConvNeXt is built entirely from standard ConvNet modules.

3.4. Experimental Results

In order to verify the superiority of the change-detection structure adopted in this study, this study re-implemented, as shown in Ref. [38], early fusion (FC-EF), mid-term conc (Fc-siam-conc), and subtract fusion (Fc-siam-diff) as F1, F2, and F3, respectively, as shown in Table 1. We also designed the late-term fusion (Fc-siam-last-conc) strategy, F4. The comparison of the MIoU accuracy results of the flooded building road change-detection results on the SpaceNet8 dataset shows that Fc-siam-conc, at 60.56% is better than all other structures, exceeding FC-EF, Fc-siam-diff, and Fc-siam-last-conc by 14.32%, 4.78%, and

1.31%, respectively. The mid-term conc fusion strategy performs outstandingly in the change detection of flooded building roads.

Table 1. Comparison of the monitoring accuracy of flooded buildings and roads with different common change-detection structures.

	Model	Background	Building			Road			MIoU	F1_Score
			Building	Flooded Building	MIoU	Road	Flooded Road	MIoU		
F1	FC-EF	97.99	51.25	15.34	33.29	40.60	26.02	33.31	46.24	58.76
F2	Fc-siam-conc	97.97	67.78	**57.39**	**62.58**	44.36	35.31	39.83	**60.56**	**71.82**
F3	Fc-siam-diff	97.57	52.39	54.17	53.28	42.28	32.50	37.39	55.78	66.60
F4	Fc-siam-last-conc	97.98	**70.19**	42.25	56.22	**47.98**	**37.84**	42.86	59.25	70.62

Figure 6 shows the experimental results of different change-detection structures. The mid-term fusion method performs better than the early fusion and late fusion methods in terms of the missed detection of flooded buildings (green graphics) and falsely detected pixels (orange graphics). (Figure 6a,d) shows that mid-term fusion has fewer missed detections of normal buildings than early fusion. Although the normal buildings are not as good as the late fusion in terms of indicators, the missed detections of normal buildings are significantly fewer (Figure 6c). In areas where there are both flooded buildings and normal buildings (roads), the false positive rates (orange graphics) of the three fusion methods are high, and the number of missed pixels on normal roads is also high (Figure 6b).

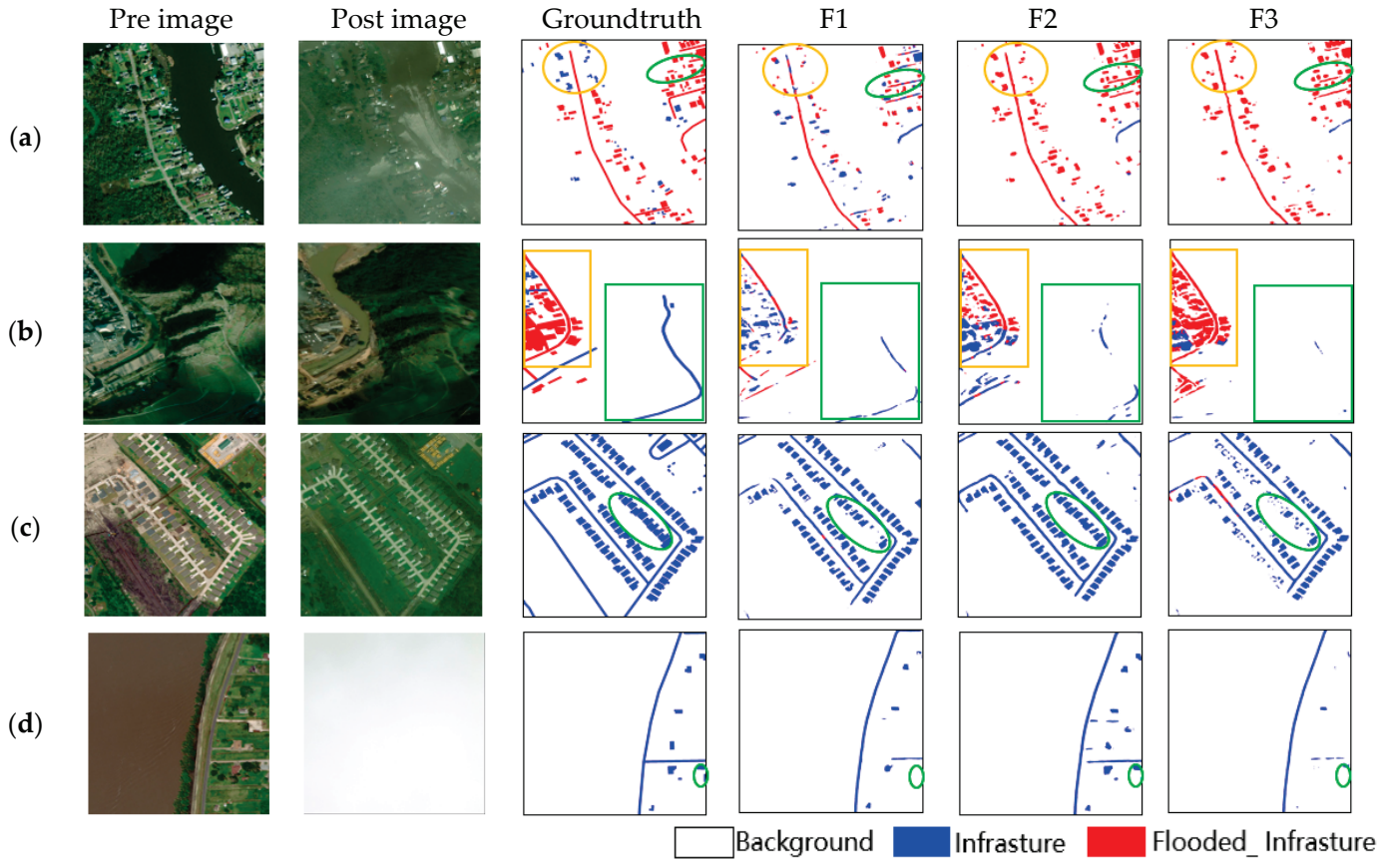


Figure 6. Comparison of monitoring results of flooded buildings and roads with different common change-detection structures. (**a**–**d**) are example figures of flooded building roads. Compared to the groundtruth, the orange shapes are a false detection, and the green shapes are a missed detection.

In order to evaluate the flooded building road change-detection model FloodedCDNet, we used EfficientNet-b5 and ResNeSt50 as a backbone to build a model with the same dual-temporal fusion structure, denoted as M2 and M3, as shown in Table 2. In the study, FC-siam-conc, ChangeFormer, and SNUNet were implemented as M1, M4, and M5. All models used the same parameters to ensure a fair comparison. The table lists the IoU accuracy of different models on the SpaceNet8 dataset. FloodedCDNet outperforms other methods in MIoU, exceeding FC-siam-conc, efficientnet-b5, ResNeSt50, ChangeFormer, and SNUNet by 7.32%, 6.73%, 6.16%, 10.02%, and 16.15%, respectively. The models M2, M3, and M6 constructed in this study outperform other change-detection models. The performance of the tranformer-based ChangeFormer is inferior to its CNN counterpart.

Table 2. Comparison of the monitoring accuracy of flooded buildings and roads using different deep learning models.

	Model	Background	Building			Road			MIoU	F1_Score
			Building	Flooded building	MIoU	Road	Flooded Road	MIoU		
M1	Fc-siam-conc	**97.97**	**67.78**	**57.39**	**62.58**	44.36	35.31	39.83	60.56	71.82
M2	FloodedCDNet$_{\text{Efficient Net-b5}}$	97.97	67.38	55.38	61.38	49.86	33.66	41.76	60.85	69.27
M3	FloodedCDNet$_{\text{ResNest50}}$	97.95	66.22	52.73	59.47	47.61	42.60	45.10	61.42	70.83
M4	changeformer	97.81	59.56	47.43	53.49	38.42	44.62	41.52	57.56	67.39
M5	Snunet	97.13	46.32	51.69	49.00	35.09	26.92	31.00	51.43	64.84
M6	FloodedCDNet$_{\text{ConvNeXt-tiny}}$	98.16	68.98	50.29	59.63	**54.32**	**49.44**	**51.88**	**64.24**	**73.65**

We conducted ablation experiments on the SpaceNet8 dataset, including the proposed modules' data augmentation, multi-scale feature fusion, and deep supervised learning, and compared their performance with baseline methods lacking these features. The table below lists the accuracies of the detection results of flooded building and road changes, indicating that the integration of these modules improves the accuracy of the evaluation indicator IoU. It is worth noting that the baseline method achieves a 6.11% lower MIoU value. This means that these modules significantly improve the accuracy of the change detection of roads to flooded buildings. In particular, the addition of the data augmentation strategy has a significant impact on the accuracy of change detection; specifically, an increase of 4.78% was observed in MIoU. At the same time, after adding MSFF, the detection accuracy of flooded building road changes increased by 1.33%.

Figure 7 and Table 3 show the ablation results, where the boundary of the changed area identified by B4 is regular and has the highest consistency with the ground truth. In particular, the addition of the MSFF module enhances the recognition of normal buildings and flooded buildings (orange graphics in Figure 7a,b,d). As shown in the orange and green graphics in (Figure 7b,c), after adding deep supervision, the flooded roads and roads become more regular and show fine boundary details.

To evaluate the simplicity of MSFF and deep supervised learning plug-and-play, and their impact on change-detection results, we conducted ablation experiments on M1, M2, and M3. The table below lists the evaluation results of Fc-siam-conc, Resnest50, Efficientnet-b5, and ConvNeXt-Tiny after adding MSFF and deep supervised learning, which shows that the detection accuracy is improved when using them. Figure 8 and Table 4 show the results of embedding the MFLL module. Compared with other methods, the results obtained using the ConvNeXt feature extractor are more accurate for flooded buildings, flooded roads, normal buildings, and normal roads (Figure 8a,b,d), with lower false detection and missing detection rates, and clear shape contours. This shows that ConvNeXt-Tiny, as a feature extractor, shows stronger robustness in identifying various objects. In addition, the problem of missed detections on normal roads is more prominent (Figure 8c,d).

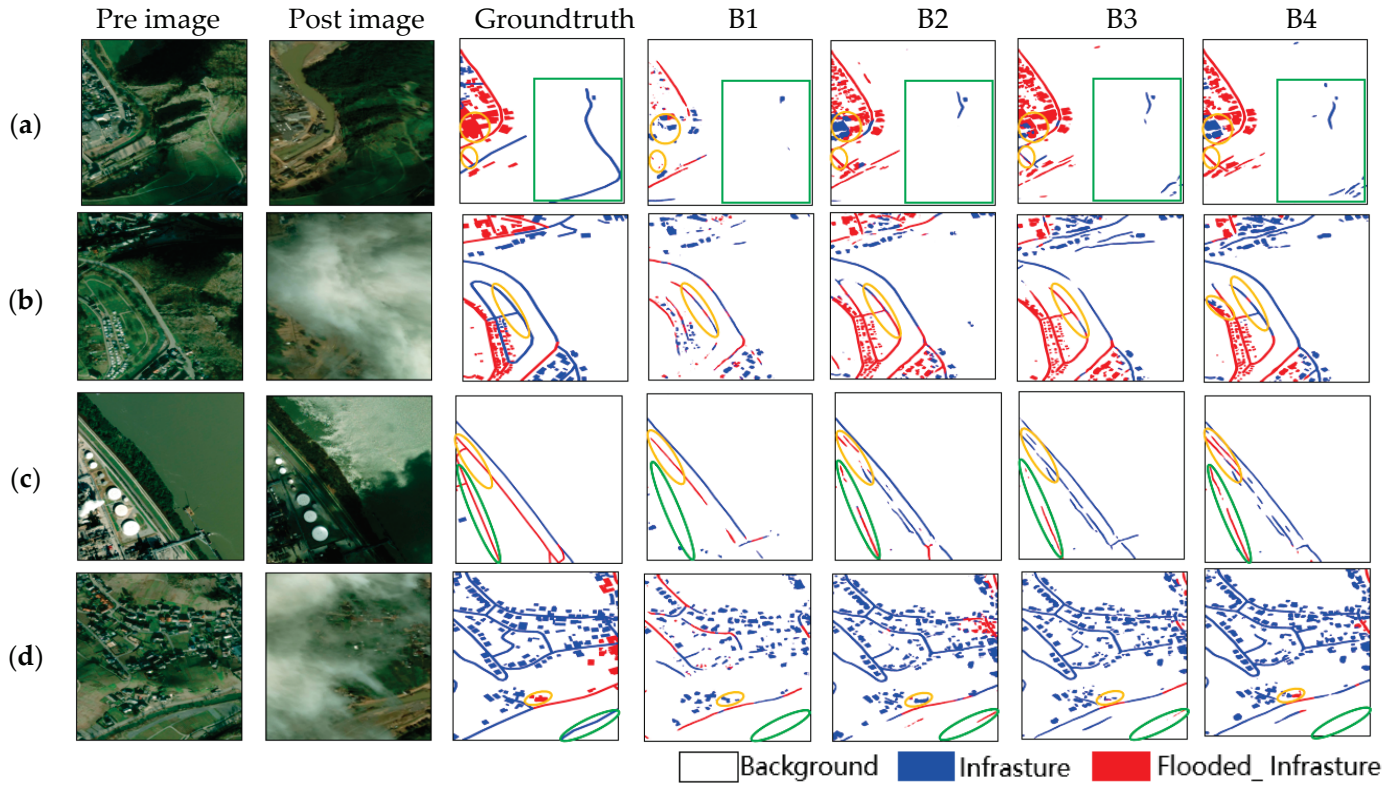


Figure 7. Comparison of ablation experiment results. (**a**–**d**) are example figures of flooded buildings and roads. Compared to the groundtruth, the orange shapes are a false detection, and the green shapes are a missed detection.

Table 3. Ablation experiment, comparison of monitoring accuracy of different modules on flooded building and road changes.

	Model	Background	Building			Road			MIoU	F1_Socre
			Building	Flooded Building	MIoU	Road	Flooded Road	MIoU		
B1	Base(FloodedCDNet$_{\text{ConvNeXt-tiny}}$)	98.16	68.98	50.29	59.63	54.32	49.44	51.88	64.24	73.65
B2	Base+data aug	98.19	73.49	63.29	68.39	56.56	46.38	51.47	67.58	76.20
B3	Base+data aug +MSFF	**98.36**	**77.50**	**65.40**	**71.45**	51.90	49.73	50.81	68.58	76.73
B4	Base +data aug +MSFF+deeply supervised	98.42	75.23	60.18	67.70	**59.40**	**51.32**	**55.36**	**68.91**	**77.42**

Table 4. Comparison of monitoring accuracy of flooded building roads with different deep learning models embedded in MSFF.

	Model	Background	Building			Road			MIoU	F1_Score
			Building	Flooded Building	MIoU	Road	Flooded Road	MIoU		
C1	Fc-siam-conc(MSFF$_{\text{all}}$)	97.97	68.05	49.51	58.78	48.43	40.34	44.38	60.86 (+0.3)	0.7136
C2	FloodedCDNetResNest50 (MSFF$_{\text{single}}$)	97.94	72.33	55.67	64.00	41.72	42.52	42.12	62.04 (+0.62)	0.7252
C3	FloodedCDNetEfficientNet-b5(MSFF$_{\text{all}}$)	97.95	67.88	50.38	59.13	46.32	46.43	46.37	61.79 (+0.94)	0.7068
C4	FloodedCDNetConvNeXt-Tiny (MSFF$_{\text{all}}$)	**98.42**	**75.23**	**60.18**	**67.70**	**59.40**	**51.32**	**55.36**	**68.91** (+1.33)	**0.7742**

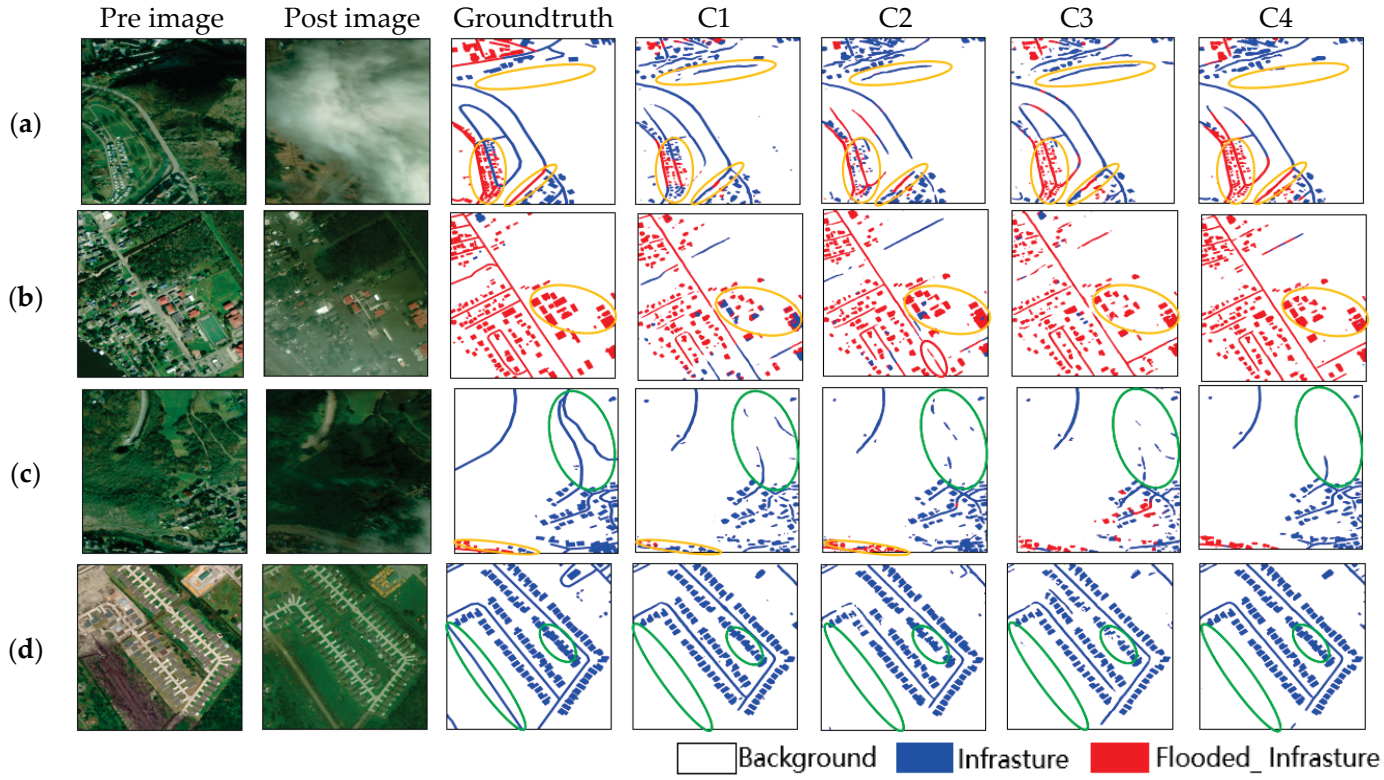


Figure 8. Comparison of monitoring results of flooded building roads embedded in different deep learning models. (**a**–**d**) are example figures of flooded building roads. Compared to the groundtruth, the orange shapes are a false detection, and the green shapes are a missed detection.

4. Discussion

This paper designs a weight-sharing encoder–decoder network FloodedCDNet and uses a mid-fusion approach to perceive the change area, obtaining semantic features from dual-temporal images based on the flood context information around the flooded buildings and the flood coverage information of the flooded roads. We use a fully convolutional ConvNeXt feature extractor inspired by the transformer, which reaches the limit of ConvNet performance and outperforms other promising CNN feature extractors. We embed a multi-attention-constrained, deeply supervised, high-dimensional and low-dimensional multi-scale feature fusion module in the feature map space, which fuses different levels of features to improve the recognition of complex changing scenes. Subsequently, deeply supervised learning was also introduced into FloodedCDNet. This method is different from simply cat-concatenating features at a specific level to form a feature map. Therefore, this study contributes to a more robust deep learning model of flooded buildings and roads, which outperforms other existing methods in terms of MIoU accuracy.

Currently, in flood disaster emergency response, the detection of flooded buildings and roads relies on pre-disaster land cover data. The information lag caused by infrequent updates may lead to misjudgment during emergency responses. As well as the scope of post-disaster flooding, most flooded buildings are not submerged by floods; therefore, the commonly used commercial methods cannot quickly detect flooded buildings. To solve this problem, this study uses a mid-fusion change-detection structure for change detection and enhances the representation of semantic features of dual-temporal building, road, and flood images. This can enhance the spatial location information of flooded buildings and roads. This study has verified that the change-detection structure is superior to several other change-detection structures in flood disaster change detection. In the future, it can be directly applied to the change-detection tasks for earthquake, landslide, mudslide, storm surge, forest, and grassland fire disaster information.

The change detection of flooded buildings and roads in very-high-resolution remote sensing images requires a more powerful feature extractor or deep learning model structure. This study focuses on the practical application of situational awareness in flood emergency response, comparing the performance of promising applications such as Fc-siam-conc, FloodedCDNet$_{\text{EfficientNet-b5}}$, FloodedCDNet$_{\text{ResNest50}}$, ChangeFormer, SNUNet, and FloodedCDNet$_{\text{ConvNeXt-Tiny}}$. The ChangeFormer based on transformer feature extraction and CNN decoding is inferior to other CNN models. Fc-siam-conc, which is commonly used in change-detection tasks, has achieved good results, but it is still not as accurate as this study's use of the network architecture search EfficientNet-b5, the modular channel attention architecture ResNeSt50, and the fully convolutional ConvNeXt-Tiny as feature extractors. FloodedCDNet$_{\text{ConvNeXt-Tiny}}$ achieved the best results for this task. This model was used as a baseline to verify the performance improvement of other modules constructed in this study. They all achieved excellent results.

Recent studies have shown that CNN-based feature extractors are not sufficient to model the distant context of ground objects, especially for complex objects [59,60]. In flood scenarios, it is not enough to simulate the flood surrounding the ground objects, which can result in poor detection results and low accuracy. By fusing high-dimensional and low-dimensional multi-scale features, MSFF enhances the feature learning of flooded building roads with complex backgrounds, shapes, and textures. This study shows that Fc-siam-conc(full), FloodedCDNet$_{\text{ResNest50}}$, FloodedCDNet$_{\text{EfficientNet-b5}}$, and FloodedCDNet$_{\text{ConvNeXt-Tiny}}$, embedded with MSFF, significantly outperform baseline methods. When embedding deeply supervised learning into FloodedCDNet$_{\text{ConvNeXt-Tiny}}$, by constructing auxiliary tasks, it further optimizes the MIoU accuracy of flooded building road change detection.

At the same time, in order to utilize the background data of buildings and roads in the disaster area before the disaster, this study designed a multi-task learning model and performed semantic change detection, including pre-disaster building road labels and flooded building road labels. After adding MSFF, its MIoU accuracy (68.19%) is lower than that of change detection (68.58%). The decline in effect may be because the remote sensing characteristics of flooded buildings after a disaster are similar to those before the disaster, and because the semantic labels of the roads of flooded buildings after the disaster are not used, causing the main branch of change detection to be ambiguous between normal and flooded buildings. In addition, an attempt was made to propose change detection assisted by boundary features, adding boundary constraints to flooded buildings and roads; the resulting accuracy was 68.16%, which is obviously lower than the change-detection accuracy. Since the boundaries of flooded building and road change-detection tasks are not regular, adding edge constraints can enhance the detection of normal buildings and roads. However, the edges have little effect on the detection accuracy in flooded building and road detection tasks. Therefore, the above two attempts are not effective in the task of detecting changes in flooded buildings and roads.

5. Conclusions

The emergency monitoring of flooded buildings and roads under the influence of natural disasters, such as floods, is crucial for emergency management departments in order to provide disaster relief. This study proposes a convolutional neural network, FloodedCDNet, for change detection in roads with flooded buildings. The FloodedCDNet architecture adopts a mid-term fusion change-detection structure. The experimental results of the common change-detection structure proposed in [42] show that the MIoU accuracy of mid-term fusion is better than that of single-branch, dual-branch-difference, and dual-branch-late fusion. It is 14.32%, 4.78% and 1.31% higher than the other three methods, respectively. The adopted feature extractor (ConvNeXt) can effectively extract the boundaries of flooded buildings and objects with clear shapes; therefore, it is very suitable for detecting building road changes of various shapes and sizes in complex flood scenarios. The performance of ConvNeXt is also better than other deep learning models, such as

FC-EF, FC-Siam-Conc, FC-Siam-Diff, EfficientNet, ResNeSt, SNUNet, ChangeFormer, etc., achieving the highest MIoU of 62.80%. This study designed a multi-attention-constrained deeply supervised high-dimensional and low-dimensional multi-scale feature fusion module to enhance the detection of flooded building roads by fusing low-dimensional detail information and high-dimensional semantic information and alleviating spatial similarity and channel redundancy through multi-attention. Compared with the baseline method, the addition of this module further improved the flooded infrastructure change-detection results, increasing the MIoU by 1.33%. This study also successfully embedded MSFF into other deep learning models. The experimental results show that the embedded MSFF performs better than the baseline model, indicating that MSFF can be used as a general multi-scale feature fusion component.

In order to further improve the usability of FloodedCDNet in flood emergency monitoring services, future work should study the effectiveness of local data (DEM, slope), aerial remote sensing images, and semi-supervised learning in flood situation perception, air-ground collaborative observation, and in the generation of high-quality pseudo labels in disaster areas.

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

Funding: This research was supported by a grant from the Key Laboratory of Emergency Satellite Engineering and Application, Ministry of Emergency Management, and the National Natural Science Foundation of China (NSFC) (no. U2243222).

Data Availability Statement: The data used in this study are available by contacting the corresponding author.

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

References

1. Kundzewicz, Z.W.; Kanae, S.; Seneviratne, S.I.; Handmer, J.; Nicholls, N.; Peduzzi, P.; Mechler, R.; Bouwer, L.M.; Arnell, N.; Mach, K.; et al. Flood risk and climate change: Global and regional perspectives. *Hydrol. Sci. J.* **2014**, *59*, 1–28. [CrossRef]
2. Shojaeian, A.; Shafizadeh-Moghadam, H.; Sharafati, A.; Shahabi, H. Extreme flash flood susceptibility mapping using a novel PCA-based model stacking approach. *Adv. Space Res.* **2024**, *74*, 5371–5382. [CrossRef]
3. Qi, W.; Ma, C.; Xu, H.; Chen, Z.; Zhao, K.; Han, H. A review on applications of urban flood models in flood mitigation strategies. *Nat. Hazards* **2021**, *108*, 31–62. [CrossRef]
4. Sepehri, M.; Malekinezhad, H.; Jahanbakhshi, F.; Ildoromi, A.R.; Chezgi, J.; Ghorbanzadeh, O.; Naghipour, E. Integration of interval rough AHP and fuzzy logic for assessment of flood prone areas at the regional scale. *Acta Geophys.* **2020**, *68*, 477–493. [CrossRef]
5. Hansch, R.; Arndt, J.; Lunga, D.; Gibb, M.; Pedelose, T.; Boedihardjo, A.; Petrie, D.; Bacastow, T.M. SpaceNet 8—The Detection of Flooded Roads and Buildings. In Proceedings of the Proceedings—2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPRW 2022, New Orleans, LA, USA, 19–22 June 2022; IEEE Computer Society: New York, NY, USA, 2022; pp. 1471–1479.
6. McCormack, T.; Campanyà, J.; Naughton, O. A methodology for mapping annual flood extent using multi-temporal Sentinel-1 imagery. *Remote Sens. Environ.* **2022**, *282*, 113273. [CrossRef]
7. Konapala, G.; Kumar, S.V.; Ahmad, S.K. Exploring Sentinel-1 and Sentinel-2 diversity for flood inundation mapping using deep learning. *ISPRS J. Photogramm. Remote Sens.* **2021**, *180*, 163–173. [CrossRef]
8. Cooner, A.J.; Shao, Y.; Campbell, J.B. Detection of Urban Damage Using Remote Sensing and Machine Learning Algorithms: Revisiting the 2010 Haiti Earthquake. *Remote Sens.* **2016**, *8*, 868. [CrossRef]
9. Brunner, D.; Lemoine, G.; Bruzzone, L. Earthquake Damage Assessment of Buildings Using VHR Optical and SAR Imagery. *IEEE Trans. Geosci. Remote Sens.* **2010**, *48*, 2403–2420. [CrossRef]
10. Wieland, M.; Fichtner, F.; Martinis, S.; Groth, S.; Krullikowski, C.; Plank, S.; Motagh, M. S1S2-Water: A Global Dataset for Semantic Segmentation of Water Bodies From Sentinel- 1 and Sentinel-2 Satellite Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 1084–1099. [CrossRef]

11. Bonafilia, D.; Tellman, B.; Anderson, T.; Issenberg, E. Sen1Floods11: A georeferenced dataset to train and test deep learning flood algorithms for sentinel-1. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 210–211.
12. Rahnemoonfar, M.; Chowdhury, T.; Sarkar, A.; Varshney, D.; Yari, M.; Murphy, R.R. FloodNet: A High Resolution Aerial Imagery Dataset for Post Flood Scene Understanding. *IEEE Access* **2021**, *9*, 89644–89654. [CrossRef]
13. Gupta, R.; Hosfelt, R.; Sajeev, S.; Patel, N.; Goodman, B.; Doshi, J.; Heim, E.; Choset, H.; Gaston, M. xBD: A Dataset for Assessing Building Damage from Satellite Imagery. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–20 June 2019; pp. 10–17. Available online: https://arxiv.org/abs/1911.09296v1 (accessed on 21 August 2024).
14. Rudner, T.G.; Rußwurm, M.; Fil, J.; Pelich, R.; Bischke, B.; Kopacková, V.; Bilinski, P. Rapid computer vision-aided disaster response via fusion of multiresolution, multisensor, and multitemporal satellite imagery. In Proceedings of the First Workshop on AI for Social Good. Neural Information Processing Systems (NIPS-2018), Montreal, QC, Canada, 3–8 December 2018; pp. 3–8.
15. Lecun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
16. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. *arXiv* **2014**, arXiv:1411.4038.
17. Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *arXiv* **2015**, arXiv:1511.00561. [CrossRef] [PubMed]
18. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv* **2015**, arXiv:1505.04597.
19. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid Scene Parsing Network. *arXiv* **2016**, arXiv:1612.01105.
20. Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. *arXiv* **2018**, arXiv:1802.02611.
21. Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. Deep High-Resolution Representation Learning for Visual Recognition. *arXiv* **2019**, arXiv:1908.07919. [CrossRef]
22. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
23. Tan, M.; Le, Q.V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *arXiv* **2019**, arXiv:1905.11946.
24. Liu, Z.; Mao, H.; Wu, C.Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A ConvNet for the 2020s. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 11966–11976. [CrossRef]
25. Lu, D.; Cheng, S.; Wang, L.; Song, S. Multi-scale feature progressive fusion network for remote sensing image change detection. *Sci. Rep.* **2022**, *12*, 11968. [CrossRef]
26. Chen, B.; Xia, M.; Huang, J. MFANet: A Multi-Level Feature Aggregation Network for Semantic Segmentation of Land Cover. *Remote Sens.* **2021**, *13*, 731. [CrossRef]
27. Long, J.; Li, M.; Wang, X.; Stein, A. Semantic change detection using a hierarchical semantic graph interaction network from high-resolution remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2024**, *211*, 318–335. [CrossRef]
28. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. *arXiv* **2019**, arXiv:1910.03151.
29. Cheng, J.; Deng, C.; Su, Y.; An, Z.; Wang, Q. Methods and datasets on semantic segmentation for Unmanned Aerial Vehicle remote sensing images: A review. *ISPRS J. Photogramm. Remote Sens.* **2024**, *211*, 1–34. [CrossRef]
30. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. *arXiv* **2018**, arXiv:1807.06521.
31. Roy, A.G.; Navab, N.; Wachinger, C. Concurrent Spatial and Channel Squeeze & Excitation in Fully Convolutional Networks. *arXiv* **2018**, arXiv:1803.02579.
32. Su, Z.; Wang, Y.; Xu, Q.; Gao, R.; Kong, Q. LodgeNet: Improved rice lodging recognition using semantic segmentation of UAV high-resolution remote sensing images. *Comput. Electron. Agric.* **2022**, *196*, 106873. [CrossRef]
33. Bo, W.; Liu, J.; Fan, X.; Tjahjadi, T.; Ye, Q.; Fu, L. BASNet: Burned Area Segmentation Network for Real-time Detection of Damage Maps in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5627913. [CrossRef]
34. Han, G.; Zhang, M.; Wu, W.; He, M.; Liu, K.; Qin, L.; Liu, X. Improved U-Net based insulator image segmentation method based on attention mechanism. *Energy Rep.* **2021**, *7*, 210–217. [CrossRef]
35. Cheng, W.; Feng, Y.; Song, L.; Wang, X. DMF2Net: Dynamic multi-level feature fusion network for heterogeneous remote sensing image change detection. *Knowl. Based Syst.* **2024**, *300*, 112159. [CrossRef]
36. Pan, J.; Cui, W.; An, X.; Huang, X.; Zhang, H.; Zhang, S.; Zhang, R.; Li, X.; Cheng, W.; Hu, Y. MapsNet: Multi-level feature constraint and fusion network for change detection. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *108*, 102676. [CrossRef]
37. Fang, S.; Li, K.; Shao, J.; Li, Z. SNUNet-CD: A Densely Connected Siamese Network for Change Detection of VHR Images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 8007805. [CrossRef]
38. Zhang, C.; Wang, L.; Cheng, S.; Li, Y. SwinSUNet: Pure Transformer Network for Remote Sensing Image Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5224713. [CrossRef]
39. Lei, T.; Geng, X.; Ning, H.; Lv, Z.; Gong, M.; Jin, Y.; Nandi, A.K. Ultralightweight Spatial-Spectral Feature Cooperation Network for Change Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 4402114. [CrossRef]

40. Song, L.; Xia, M.; Weng, L.; Lin, H.; Qian, M.; Chen, B. Axial Cross Attention Meets CNN: Bibranch Fusion Network for Change Detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 32–43. [CrossRef]
41. Zheng, Z.; Ma, A.; Zhang, L.; Zhong, Y. Change is Everywhere: Single-Temporal Supervised Object Change Detection in Remote Sensing Imagery. In Proceedings of the IEEE International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 15173–15182. [CrossRef]
42. Chen, H.; Qi, Z.; Shi, Z. Remote Sensing Image Change Detection with Transformers. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5607514. [CrossRef]
43. Zheng, Z.; Ermon, S.; Kim, D.; Zhang, L.; Zhong, Y. Changen2: Multi-Temporal Remote Sensing Generative Change Foundation Model. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, 1–17. [CrossRef]
44. Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Li, F.-F. ImageNet: A Large-Scale Hierarchical Image Database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, Miami, FL, USA, 20–25 June 2009; pp. 248–255. [CrossRef]
45. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. April 2017. Available online: http://arxiv.org/abs/1704.04861 (accessed on 21 August 2024).
46. Ba, J.L.; Kiros, J.R.; Hinton, G.E. Layer Normalization. *arXiv* **2016**, arXiv:1607.06450.
47. Hendrycks, D.; Gimpel, K. Gaussian Error Linear Units (GELUs). *arXiv* **2016**, arXiv:1606.08415.
48. Yang, M.; Yu, K.; Zhang, C.; Li, Z.; Yang, K. DenseASPP for Semantic Segmentation in Street Scenes. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3684–3692. [CrossRef]
49. Yu, F.; Wang, D.; Shelhamer, E.; Darrell, T. Deep Layer Aggregation. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2403–2412. [CrossRef]
50. Miao, L.; Li, X.; Zhou, X.; Yao, L.; Deng, Y.; Hang, T.; Zhou, Y.; Yang, H. SNUNet3+: A Full-Scale Connected Siamese Network and a Dataset for Cultivated Land Change Detection in High-Resolution Remote-Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 4400818. [CrossRef]
51. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-Excitation Networks. *arXiv* **2017**, arXiv:1709.01507.
52. Mao, A.; Mohri, M.; Zhong, Y. Cross-Entropy Loss Functions: Theoretical Analysis and Applications. *arXiv* **2023**, arXiv:2304.07288.
53. Li, X.; Sun, X.; Meng, Y.; Liang, J.; Wu, F.; Li, J. Dice Loss for Data-imbalanced NLP Tasks. *arXiv* **2019**, arXiv:1911.02855.
54. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. *arXiv* **2017**, arXiv:1708.02002.
55. Lee, C.-Y.; Xie, S.; Gallagher, P.; Zhang, Z.; Tu, Z. Deeply-Supervised Nets. *arXiv* **2014**, arXiv:1409.5185.
56. Ghorbanzadeh, O.; Shahabi, H.; Piralilou, S.T.; Crivellari, A.; la Rosa, L.E.C.; Atzberger, C.; Li, J.; Ghamisi, P. Contrastive Self-Supervised Learning for Globally Distributed Landslide Detection. *IEEE Access* **2024**, *12*, 118453–118466. [CrossRef]
57. Zhang, H.; Wu, C.; Zhang, Z.; Zhu, Y.; Lin, H.; Zhang, Z.; Sun, Y.; He, T.; Mueller, J.; Manmatha, R.; et al. ResNeSt: Split-Attention Networks. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, New Orleans, LA, USA, 19–20 June 2022; pp. 2735–2745. [CrossRef]
58. Bandara, W.G.C.; Patel, V.M. A Transformer-Based Siamese Network for Change Detection. In *International Geoscience and Remote Sensing Symposium (IGARSS)*; Institute of Electrical and Electronics Engineers Inc.: New York, NY, USA, 2022; pp. 207–210. [CrossRef]
59. Ghorbanzadeh, O.; Gholamnia, K.; Ghamisi, P. The application of ResU-net and OBIA for landslide detection from multi-temporal Sentinel-2 images. *Big Earth Data* **2023**, *7*, 961–985. [CrossRef]
60. Iqbal, U.; Perez, P.; Li, W.; Barthelemy, J. How computer vision can facilitate flood management: A systematic review. *Int. J. Disaster Risk Reduct.* **2021**, *53*, 102030. [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

DBSANet: A Dual-Branch Semantic Aggregation Network Integrating CNNs and Transformers for Landslide Detection in Remote Sensing Images

Yankui Li [1], Wu Zhu [1,2,3,*], Jing Wu [4], Ruixuan Zhang [1], Xueyong Xu [5] and Ye Zhou [5]

[1] School of Geological Engineering and Geomatics, Chang'an University, Xi'an 710054, China; 2022126062@chd.edu.cn (Y.L.); 2023026003@chd.edu.cn (R.Z.)
[2] Key Laboratory of Ecological Geology and Disaster Prevention, Ministry of Natural Resources, Xi'an 710054, China
[3] Big Data Center for Geosciences and Satellites, Xi'an 710054, China
[4] Qinghai Institute of Geo-Environment Monitoring, Xining 810007, China; endless5wj@163.com
[5] North Information Control Research Academy Group Co., Ltd., Nanjing 211153, China; xxyyeah@163.com (X.X.); leaf430ahip@126.com (Y.Z.)
* Correspondence: zhuwu@chd.edu.cn

Abstract: Deep learning-based semantic segmentation algorithms have proven effective in landslide detection. For the past decade, convolutional neural networks (CNNs) have been the prevailing approach for semantic segmentation. Nevertheless, the intrinsic limitations of convolutional operations hinder the acquisition of global contextual information. Recently, Transformers have garnered attention for their exceptional global modeling capabilities. This study proposes a dual-branch semantic aggregation network (DBSANet) by integrating ResNet and a Swin Transformer. A Feature Fusion Module (FFM) is designed to effectively integrate semantic information extracted from the ResNet and Swin Transformer branches. Considering the significant semantic gap between the encoder and decoder, a Spatial Gate Attention Module (SGAM) is used to suppress the noise from the decoder feature maps during decoding and guides the encoder feature maps based on its output, thereby reducing the semantic gap during the fusion of low-level and high-level semantic information. The DBSANet model demonstrated superior performance compared to existing models such as UNet, Deeplabv3+, ResUNet, SwinUNet, TransUNet, TransFuse, and UNetFormer on the Bijie and Luding datasets, achieving IoU values of 77.12% and 75.23%, respectively, with average improvements of 4.91% and 2.96%. This study introduces a novel perspective for landslide detection based on remote sensing images, focusing on how to effectively integrate the strengths of CNNs and Transformers for their application in landslide detection. Furthermore, it offers technical support for the application of hybrid models in landslide detection.

Keywords: landslide detection; remote sensing; convolutional neural network; transformer; feature fusion; semantic gap

1. Introduction

Landslides occur due to gravitational or other factors, causing soil or rock masses to slide along vulnerable surfaces of slopes, thereby posing a significant threat to human lives, the socio-economy, and the natural environment [1]. Rapid and accurate landslide detection is vital for post-disaster emergency response and comprehensive damage assessment [2]. Traditional manual field survey methods have significantly limited effective applications

in post-disaster large-scale landslide detection tasks due to their substantial cost and relative inefficiency [3].

The rapid advancements in aerospace and sensor technologies have provided researchers with an expanding repository of high-quality remote sensing images, thereby enabling the non-contact detection of landslides using these images [4]. Currently, landslide detection based on remote sensing images primarily encompasses four methodologies: visual interpretation, pixel-based, object-based, and deep learning methods. Visual interpretation requires rich experience and is greatly influenced by human subjectivity [5]. The pixel-based approach designates each pixel as the classification unit and performs classification by extracting features from individual pixels. But it overlooks the contextual information from neighboring pixels and is susceptible to interference from similar spectral information features [6]. The object-based method groups neighboring pixels with similar attributes into the homogeneous patches. Compared to the pixel-based method, these approaches take into account spatial, spectral, and shape features [7]. However, achieving optimal segmentation often requires iterative parameter adjustments, making the classification process more complex and not suitable for large-scale data processing. Over the past few years, deep learning techniques have gained widespread application in extracting information from remote sensing imagery, attributable to their robust feature extraction capabilities [8]. Now, landslide detection approaches utilizing deep learning are categorized into object detection [9,10] and semantic segmentation [11,12] techniques. Among them, semantic segmentation is a pixel-level categorization that distinguishes between the foreground and background by determining the category to which each pixel belongs.

The rapid development of convolutional neural networks (CNNs) has significantly propelled research in semantic segmentation of remote sensing images [13]. Long et al. [14] introduced the Fully Convolutional Network (FCN), which pioneered end-to-end pixel-level image segmentation. Subsequently, numerous scholars have developed a series of enhanced models based on this architecture for landslide detection [15,16]. For example, Li and Guo [17] utilized MobileNetV2 for extracting landslide features, effectively enhancing the detection speed. Qi et al. [18] significantly improved the accuracy of landslide detection by combining the residual structure with the UNet model to design Res-Unet. Although CNN-based models can achieve good segmentation performance, they depend on convolutional kernels for feature extraction, which limits their ability to capture global semantic information in remote sensing images due to fixed receptive field sizes. Nevertheless, capturing global contextual information is essential for landslide detection tasks. Researchers commonly adopt two methods to address this issue. One approach involves modifying the convolution operation, such as employing large convolutional kernels, dilated convolutions, or employing feature pyramid pooling, thereby expanding the receptive field. For instance, Chen et al. [19] proposed DeepLabV3+, which integrates Atrous Spatial Pyramid Pooling (ASPP), leveraging dilated convolutions to substantially expand the receptive field while preserving the parameter count. Xia et al. [20] introduced the ASPP module to improve the model's capability to capture multi-scale contextual information from high-resolution remote sensing images. The alternative approach focused on optimizing the weight distribution across channel or spatial dimensions through the integration of attention mechanisms within CNN architectures, thereby emphasizing information pertinent to landslides. Generally, attention mechanisms are broadly categorized into three types: channel attention, spatial attention, and mixed attention. Chen et al. [21] integrated channel attention with Unet for landslide detection in Sentinel-2A images. However, both of these approaches concentrate on optimizing the CNN structure and do not fully address the inherent limitations of convolution operation. Meanwhile, CNNs based on the encoder–decoder structure overlook the inherent semantic disparity between shallow feature maps

in the encoder and deeper counterparts in the decoder. The semantic levels of image features captured by shallow and deep feature maps are inconsistent due to variations in the number of related operations. However, most current semantic segmentation models only adopt simple skip connections to directly concatenate shallow and deep feature maps, thus overlooking the aforementioned semantic disparity, thereby impacting the segmentation model's overall performance. To address this issue, researchers have undertaken numerous efforts. For instance, Zhou et al. [22] introduced UNet++, which effectively linked feature maps from corresponding stages of the encoder and decoder through a sequence of nested dense skip connections. Pang et al. [23] proposed SENet, aiming to enhance the representation of shallow feature maps through depth-separable Atrous Spatial Pyramid Pooling, thereby narrowing the semantic gap.

A Transformer [24] was initially applied in the field of Natural Language Processing (NLP) and demonstrated exceptional performance within this domain. Its self-attention mechanism can capture the long-distance dependencies between pixels, which can better identify the shape and structure of targets. Based on this, researchers have increasingly integrated Transformers into image semantic segmentation, thus providing a solution to the problem that semantic segmentation methods based on CNNs struggle to globally model image information. Currently, Transformer models used for image semantic segmentation tasks can generally be classified into two types. The first category refers to a pure Transformer architecture solely relying on the self-attention mechanism for feature extraction and pixel classification. For instance, Cao et al. [25] constructed SwinUNet, a symmetric encoder–decoder structure based on Swin Transformer Blocks. The alternate category combines CNNs and Transformers in different ways, aiming to leverage the strengths of each model type to obtain a more comprehensive representation of semantic features [26–28]. These works fully demonstrate the suitability of the Transformer models in performing semantic segmentation on images. Given the powerful ability of Transformers in global context modeling, researchers have advanced the application of the Transformer architecture to landslide detection using remote sensing imagery. Lv et al. [29] integrated enhanced shape information with the Vision Transformer (ViT) model to perform landslide detection in optical imagery. Tang et al. [30] adopted the SegFormer for coseismic landslide detection, complemented by image processing techniques to eliminate spurious gaps in segmentation outcomes. Although the Transformer model excels at capturing the global context, it faces challenges in extracting fine-grained information from images [31,32]. In response to this challenge, researchers have explored the integration of CNNs and Transformers for landslide detection. For example, Li et al. [33] innovatively applied a modified version of VGG-16 integrated with a Multi-scale Lightweight Transformer (MLT) module to precisely identify landslide features in high-resolution remote sensing images. Similarly, Yang et al. [34] employed an alternative approach by utilizing the ResUNet and ViT models as robust encoders, augmented with the CBAM, to improve the efficacy of landslide detection. Notably, these methods all utilize a tandem architecture, wherein images are initially preprocessed and resized by CNNs before being input into the Transformer component for subsequent processing. While these methods have shown promising potential for landslide detection, research on hybrid models that combine CNNs and Transformers remains in the early stages of exploration.

Therefore, exploring effective approaches to integrating the strengths of CNNs and Transformers, while concurrently bridging the semantic gap between encoder and decoder layer feature mappings, is worthwhile for obtaining better feature representations and applying them to landslide detection. In this study, we propose a new semantic segmentation model named DBSANet for landslide detection. DBSANet employs a dual-branch parallel structure of ResNet and Swin Transformer to acquire local and global feature information

from remote sensing images. To fully leverage the performance of the hybrid network, we propose an innovative Feature Fusion Module (FFM) designed to aggregate the local feature information captured by ResNet with the global context information obtained from Swin Transformer. Additionally, to bridge the semantic gap between the encoder and decoder, we design a Spatial Gated Attention Module (SGAM) and embed it into the skip connection part.

2. Materials and Methods

2.1. Dataset

To evaluate the performance of the proposed DBSANet model, we have selected the Bijie [35] and Luding datasets for the experiments conducted in this study (Figure 1). The Bijie dataset, recognized as a prominent public resource for landslide analysis, comprises 770 landslide images of diverse sizes, sourced from TripleSat satellite imagery. The Luding dataset is studied in a specific case: on 5 September 2022, a magnitude 6.8 earthquake occurred in Luding County, located within the Ganzi Tibetan Autonomous Prefecture of Sichuan Province (Figure 1), triggering a series of landslides. Utilizing the GF-2 satellite data (orbit number 43565), acquired on 10 September 2022, we meticulously constructed a specialized dataset focused on landslide hazards. The production process mainly includes radiometric calibration, orthorectification, and band fusion, resulting in high-resolution multispectral images. Based on the acquired images, visual interpretation was conducted, successfully identifying and annotating 283 landslide images. To further enrich the Luding dataset and improve its generalization ability, we employed data augmentation techniques such as horizontal and vertical flipping, resulting in an expanded dataset totaling 849 landslide images. During the experimental phase, the images from the Bijie and Luding datasets were uniformly resized to 256 × 256 pixels and randomly divided into training, validation, and testing sets in a ratio of 7:2:1.

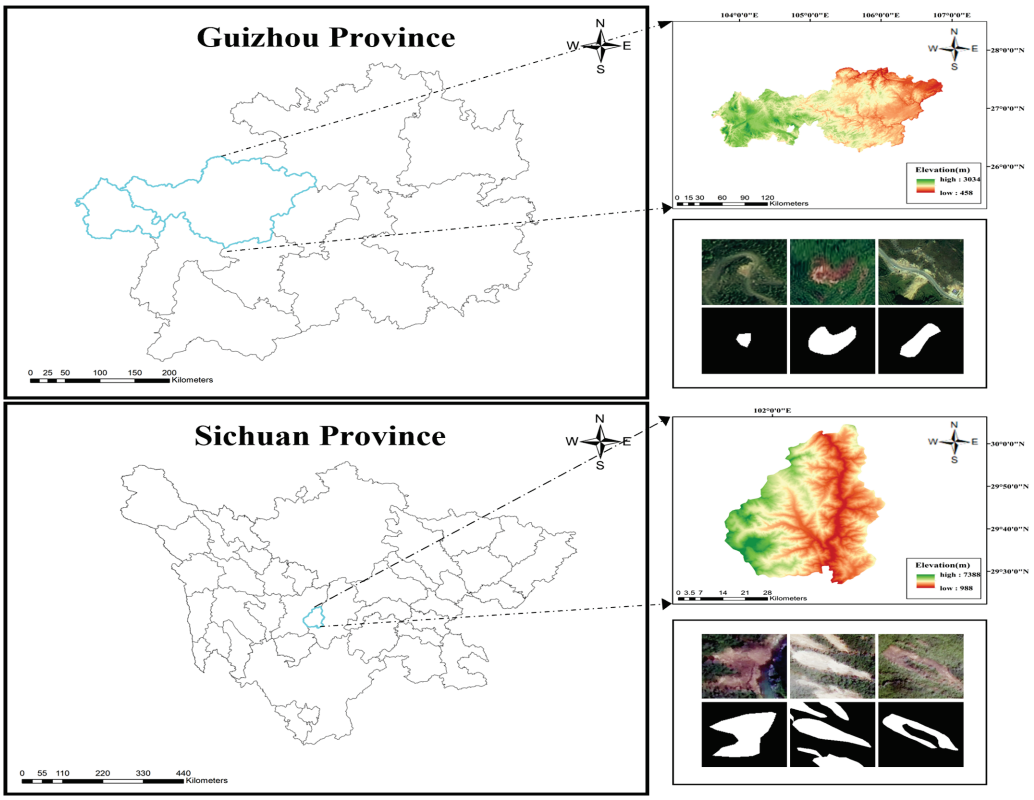


Figure 1. The coverage of the collected dataset.

2.2. Method

2.2.1. Overall Structure of the DBSANet

Figure 2 illustrates the overarching architecture of DBSANet, featuring a classical U-shaped encoder–decoder design. In the encoder part, DBSANet employs a dual-branch parallel design with the CNN and Transformer. Specifically, the CNN branch uses ResNet-18 as the backbone, while the Transformer branch adopts Swin Transformer (Swin-T). The semantic features extracted from these two branches are effectively fused through an FFM, which fully combines the advantages of the two model structures. In the decoder part, DBSANet incrementally restores the image resolution through the use of transpose convolutions. Residual blocks are additionally introduced in the decoder to augment feature extraction capabilities and reduce channel dimensions. Moreover, DBSANet integrates an SGAM within its skip connections, with the objective of reducing the semantic disparity between the encoder and decoder features. Considering that directly upsampling the decoder-derived features by a factor of 4 to recover the original image resolution may result in notable spatial detail loss, DBSANet initially produces two feature maps of varying scales-$H \times W \times C/2$ and $H/2 \times W/2 \times C/2$—via residual blocks and max pooling operations, where H and W denote the height and width of the image, respectively, while C refers to the number of channels within the image. This approach allows the decoder to incrementally upsample features by a factor of 2 to achieve the original image resolution, and subsequently, the resultant feature map undergoes a 1×1 convolution for obtaining the final segmentation outcome.

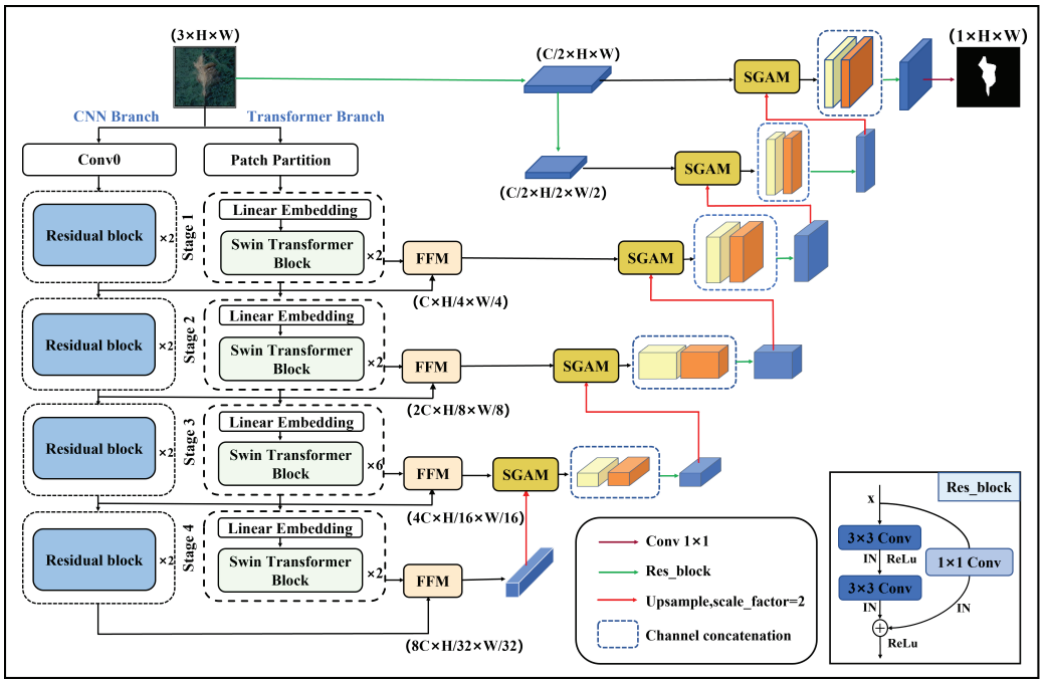


Figure 2. Structure of the proposed DBSANet.

2.2.2. CNN-Based Encoder

In semantic segmentation tasks, deeper networks can capture more image details, thereby improving the classification accuracy of foreground and background pixels. However, studies have found that increasing the number of network layers may result in model "degradation". To solve this issue, He et al. [15] introduced the ResNet, which incorporated an innovative residual connection design, enabling the construction of even deeper network models while maintaining good training stability and performance. While alternative variants of ResNet may provide some degree of performance improvement, they also involve

additional computational resources. In this study, we carefully consider the trade-offs between efficiency and accuracy, with a pronounced preference for the latter. Consequently, the CNN branch of DBSANet chooses ResNet-18 as the backbone for feature extraction. In contrast to deeper ResNet configurations, ResNet-18 boasts a comparatively straightforward network architecture, leading to a more stable training process. In processing the input remote sensing image, Conv0 initially adjusts its shape to $H/4 \times W/4 \times C$, followed by four stages of feature extraction, each incorporating a specific number of residual blocks. The output of the ith stage is denoted as R_i, which has the shape shown in Figure 2, where $C = 64,\ i = 1, 2, 3, 4$.

2.2.3. Swin Transformer-Based Encoder

Swin Transformer [36] is a deep neural network based on the Transformer architecture, featuring a hierarchical network structure and windowing strategy that adeptly capture global information while markedly reducing the computational complexity. Swin Transformer comes in four variants—Swin-T, Swin-S, Swin-B, and Swin-L—distinguished by their parameter counts. Given Swin-T's optimal balance of performance and efficiency among Swin Transformer variants, this study's Transformer branch employs Swin-T for feature extraction, as illustrated in Figure 2. Firstly, the input remote sensing image is segmented into independent patches via patch partitioning, with each patch having a dimension of 48, and a total of $H/4 \times W/4$ patches. Subsequently, the image undergoes four successive stages of feature extraction, with each stage's output denoted as T_i, characterized by its shape, which has the shape shown in Figure 2, where $C = 64, i = 1, 2, 3, 4$. In Stage 1, the two-dimensional sequential features undergo transformation into one-dimensional sequential features via Linear Embedding. Subsequently, the channel count is projected to C before entering the Swin Transformer Block. The subsequent three stages consist of Patch Merging and the Swin Transformer Block. Patch Merging is employed to halve the dimensions H and W of the image, while simultaneously doubling the channel dimension. At the heart of the Swin Transformer lies the Swin Transformer Block, pivotal for its design. It substitutes the Multi-head Self-Attention (MSA) of conventional Transformers with the Window-based Multi-head Self-Attention (W-MSA) and Shifted Window-based Multi-head Self-Attention (SW-MSA) modules. These two configurations manifest sequentially and in pairs, as presented in Figure 3. The Swin Transformer significantly mitigates the computational complexity by partitioning the feature map into numerous non-overlapping windows using the Windows Multi-Head Self-Attention (W-MSA) mechanism, with self-attention computations confined to these windows. Concurrently, to facilitate cross-window information exchange, the SW-MSA is introduced to sequentially link the windows, fostering efficient information interaction. The specific computation process and formulas are delineated as follows:

$$\hat{Z}^l = W\text{-}MSA\left(LN\left(Z^{l-1}\right)\right) + Z^{l-1} \tag{1}$$

$$Z^l = MLP\left(LN\left(\hat{Z}^l\right)\right) + \hat{Z}^l \tag{2}$$

$$\hat{Z}^{l+1} = SW\text{-}MSA\left(LN\left(Z^l\right)\right) + Z^l \tag{3}$$

$$Z^{l+1} = MLP\left(LN\left(\hat{Z}^{l+1}\right)\right) + \hat{Z}^{l+1} \tag{4}$$

where $\hat{Z}^l$ and $\hat{Z}^{l+1}$ represent the outputs of the W-MSA and SW-MSA modules, respectively; Z^l and Z^{l+1} represent the outputs of the MLP in two consecutive Swin Transformer blocks, respectively.

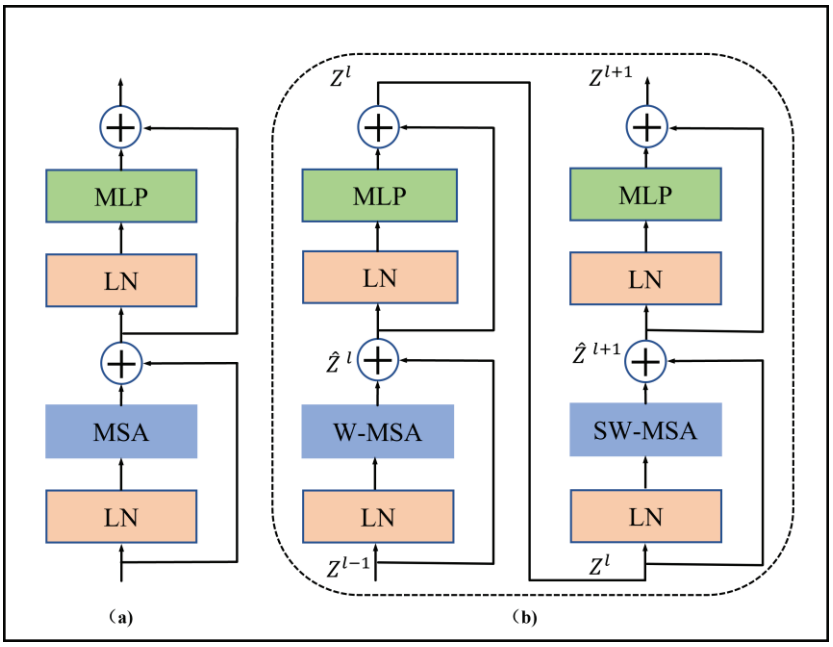


Figure 3. (**a**) Transformer block. (**b**) Swin Transformer Block.

2.2.4. Feature Fusion Module

Previous research has shown that CNNs can better capture the local context information in images. However, the fixed receptive field size of their convolutional kernels poses challenges in effectively modeling global contextual information. Transformers can better model long-distance dependencies through their self-attention mechanism, but there is a problem of losing detailed information. For a more intuitive grasp of the disparities in feature extraction between Swin-T and ResNet-18, we visualized the feature maps extracted by each of them, as shown in Figure 4. The figure illustrates discernible distinctions in the extracted features between the two models. Specifically, the initial layers of ResNet-18's feature extraction backbone primarily capture elementary semantic features like points, lines, and local textures from images. With increasing network depth, it gradually extracts high-level semantic information, showcasing an evolutionary transition from specific to abstract characteristics. In contrast, the feature extraction process of Swin-T appears more direct and global in vision. Its superficial features already manifest distinct "abstract" attributes, whereas deeper layers display a notable "clustering" phenomenon, indicating the closeness and similarity of relationships between pixels. Simply adding the features extracted by both models could result in redundancy or conflicts due to feature inconsistencies, potentially impacting the overall accuracy of landslide detection.

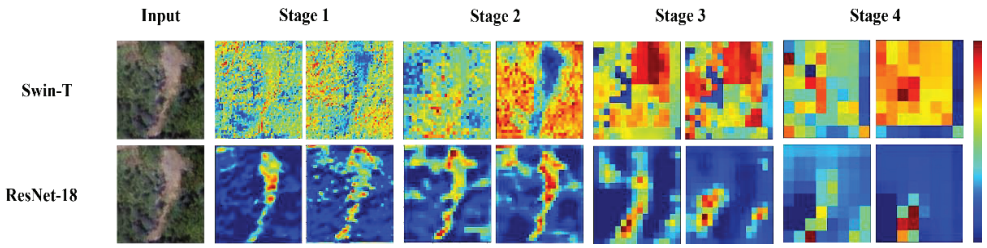


Figure 4. Example of feature map visualization in the Luding dataset, where Stage 1, Stage 2, Stage 3, and Stage 4 correspond to the four stages of feature extraction for the four backbone encoders (any two output channels). The feature maps are depicted using COLORMAP_JET pseudocolors to enhance visual differentiation. The COLORBAR is on the rightmost side of the figure, and the pixel values correspond to the colors increasing from bottom to top.

Inspired by style transfer [37] and spatial feature transformation [38], we design a Feature Fusion Module, with its specific structure shown in Figure 5. As illustrated in Figure 5, the sequential features (B, N, and C) generated by Swin-T are initially restructured into a spatial format (B, C, H, and W), thereby ensuring consistency with the spatial features derived from the CNN branch. Instance normalization (IN) is additionally employed to alleviate style variations (e.g., contrast) within individual samples. Concurrently, the feature maps from the CNN branch undergo initial processing through two distinct 3×3 convolutional layers, a procedure designed to yield a critical pair of affine transformation parameters (γ and β). Subsequently, this pair of parameters is then used to precisely adjust the scale and offset of the feature maps from the Transformer branch. By performing the above, the features extracted from both branches can be more effectively integrated, thereby preserving their inherent strengths. Meanwhile, to further augment the network's representational prowess, the SRM ResBlock is integrated into the FFM. The Style-based Recalibration Module (SRM), a pioneering module introduced by Lee et al. [39] cleverly integrates style transfer into the channel attention mechanism, thereby enabling CNNs to comprehensively capture stylistic features in images. In this study, we integrate the SRM with ResBlock to strengthen the representation of landslide features and enhance the model's generalization capabilities. This integration enables the network to adeptly adjust to feature variations across diverse landslide scenarios, thereby enhancing the prediction accuracy and reliability. The specific calculation process is described in the following formula:

$$F^i = SRM\ ResBlock\big(\big(Conv(R_i) \times IN(T_i)\big) + Conv(R_i)\big) \tag{5}$$

where F_i denotes the feature map after fusion in four stages, with $i = 1, 2, 3, 4$. R_i and T_i represent the outputs from the CNN and Transformer branches in four stages, respectively. IN denotes the instance normalization operation.

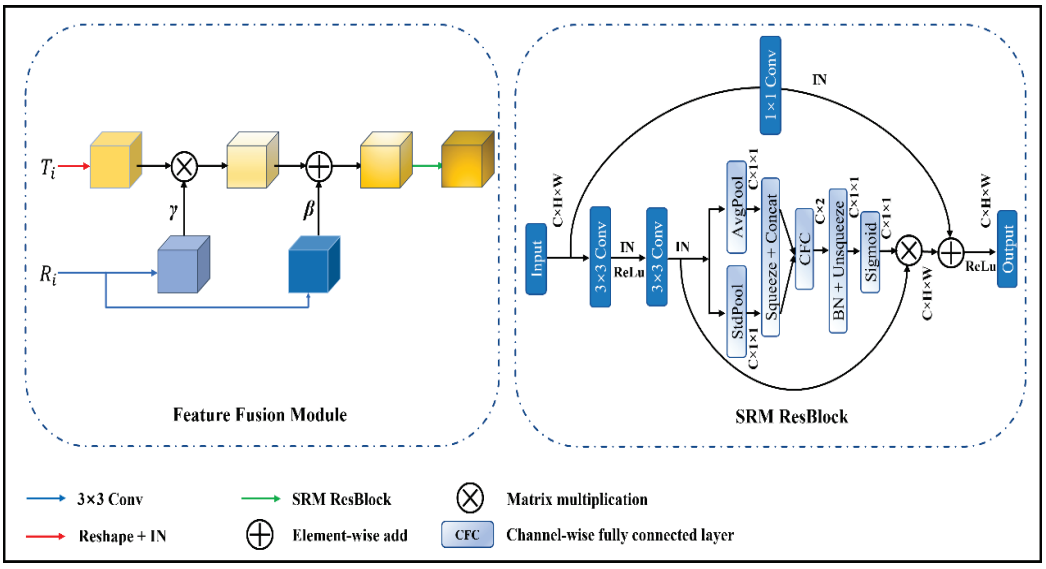


Figure 5. Structure of the proposed FFM.

2.2.5. Spatial Gate Attention Module

In the semantic segmentation task, the encoder–decoder architecture is a commonly utilized model structure in which the encoder extracts multi-level feature representations from the input image, while the decoder restores the feature map output by the encoder to the original image resolution and generates pixel-level classification results. However, research has found a semantic gap between the encoder's and decoder's output features. Shallow feature maps, for instance, often contain intricate details of the image, including

boundary shapes and texture complexities—critical cues for image recognition. Conversely, deep feature maps encapsulate richer semantic information of the image. These two types of feature maps play indispensable roles in semantic segmentation tasks, and their complementarity is essential for enhancing the model's performance. Traditional methods typically concatenate shallow and deep feature maps along the channel dimensions through skip connections. Despite enhancing segmentation performance to a certain degree, its efficacy remains constrained by overlooking the semantic hierarchy discrepancies among feature maps at varying levels. Therefore, how to effectively fuse these feature maps and minimize the semantic gap between them is crucial for enhancing the performance of semantic segmentation models.

To effectively bridge the semantic gap between the encoder and decoder, we design a Spatial Gated Attention Module, with its specific structure depicted in Figure 6. Firstly, the SGAM transmits decoder feature maps to the spatial attention unit, which adaptively weights crucial regions to enhance their prominence, thereby enabling the model to better focus on areas related to landslides. Subsequently, the encoder feature maps are guided by the obtained outputs, aiming to model the correlation between the shallow feature maps and the deeper feature maps, and leveraging the learned intrinsic relationship between them to reweight the shallow feature maps. Ultimately, after the elaborate processing of the SGAM, feature maps from both the encoder and decoder undergo reweighting before being outputted. Through the above steps, the semantic gap between the shallow and deep feature maps is effectively minimized, thereby significantly enhancing the model's performance. To make this module more efficient in practical use, the SGAM incorporates Depthwise Separable Convolution (DSC) [40]. DSC decomposes conventional convolution operations into depthwise and pointwise convolutions, thereby substantially lowering the computational overhead and model parameters without compromising performance. The above process can be represented by Equations (6)–(9).

$$F_d' = Sigmoid\Big(Conv\big(Concat(max(F_d), mean(F_d))\big)\Big) \times F_d \tag{6}$$

$$N_1 = Sigmoid\Big(Conv\big((F_d' + F_i)\big)\Big) \times F_i \tag{7}$$

$$N_2 = DSC\Big(Sigmoid\big(DSC(F_d')\big) \times Tanh\big(DSC(F_d')\big)\Big) \tag{8}$$

$$F_i' = Relu(N_1 + N_2) \tag{9}$$

where F_d denotes the feature map from the decoder, F_d' represents the output result after passing through the spatial attention unit, F_i denotes the output of the FFM, F_i' represents the reweighted encoder feature map, *DSC* stands for Depthwise Separable Convolution, *max* and *mean* represent the maximum channel compression and average channel compression, respectively, while *Concat* refers to concatenating the outputs along the channel dimensions.

2.3. Evaluation Metrics

In this study, three commonly employed semantic segmentation evaluation metrics—accuracy, Intersection over Union (IoU), and F1_score—are chosen to thoroughly and impartially assess the classification outcomes. The IoU quantifies the degree of overlap between the landslide area predicted by the model and the actual landslide area. The accuracy reflects the model's overall performance in classifying the entire dataset, demonstrating its ability to effectively differentiate between the landslide and non-landslide areas.

The F1-score demonstrates the model's ability to maintain a balanced performance between precision and recall. The calculation formulas are given as follows [41]:

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

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

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

$$F1\_Score = 2\frac{precision \times recall}{precision + recall} \tag{13}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{14}$$

where *TP* denotes the number of pixels accurately identified as landslides; *TN* denotes the number of pixels accurately identified as background; *FP* denotes the number of pixels where the background is incorrectly recognized as a landslide; and *FN* denotes the number of pixels where the landslide is incorrectly recognized as background.

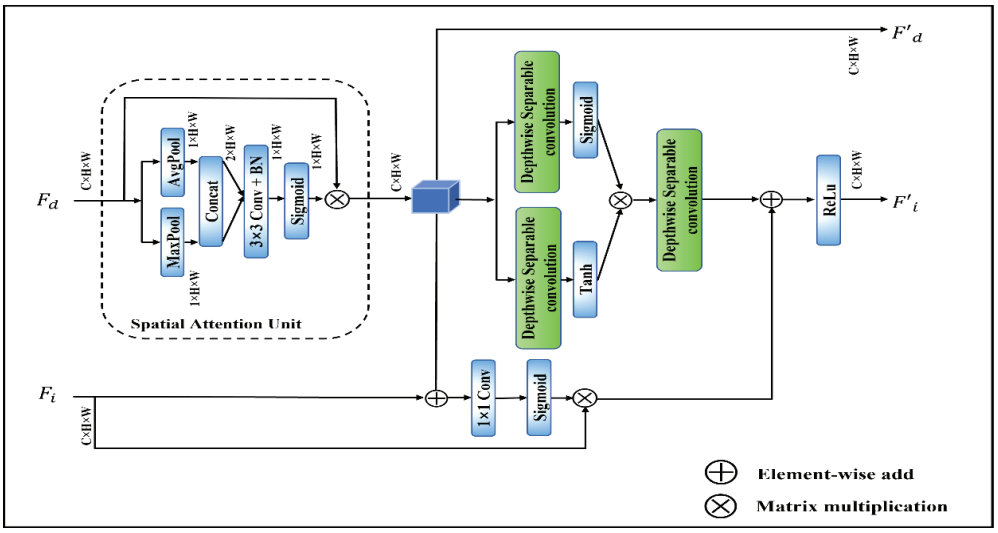


Figure 6. Structure of the proposed SGAM.

2.4. Loss Function

During the training process, the choice of loss function plays a crucial role in the optimization of the model. In the data preprocessing phase, we performed a comprehensive statistical analysis of the number of non-landslide and landslide pixel values in the corresponding landslide labels. The results indicated that the proportion of landslide pixels in the images is relatively small, representing a typical class imbalance problem. Given the relatively low proportion of positive samples in the landslide dataset compared to negative samples, the network employs a joint loss function during training to mitigate potential bias towards categories with fewer instances. It consists of weighted cross-entropy loss [42] and dice loss [43]. The weighted cross-entropy loss function is based on the cross-entropy loss function, which assigns different weights to the landslide and background. This is so that the model focuses more on categories with larger weights during the training process, thereby alleviating the problem of class imbalance to some extent. The dice loss function quantifies the overlap between the predicted outcomes and true labels, which can more

effectively address this imbalance and enhance the model's focus on foreground pixels throughout training. The joint loss function can be expressed as follows:

$$L_{total} = (1 - \alpha)L_{wce} + \alpha L_{dice} \tag{15}$$

where α is a hyperparameter, which is used to assign weights to the two loss functions. The calculation formulas for L_{wce} and L_{dice} are as follows:

$$L_{wce} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C}\alpha_c\gamma_n log\frac{exp(\gamma'_n)}{\sum_{j=1}^{C}exp(\gamma'_n)} \tag{16}$$

$$L_{dice} = 1 - \frac{2\left(\sum_{n=1}^{N}\gamma_n\gamma'_n\right)}{\sum_{n=1}^{N}\gamma_n + \sum_{n=1}^{N}\gamma'_n} \tag{17}$$

where N and C denote the number of samples and categories, respectively; α_c refers to the weight of category c; γ'_n and γ_n denote the network prediction segmentation result and ground truth, respectively.

2.5. Implementation Details

The study's experiments were conducted using the Windows 11 operating system and leveraging an NVIDIA GeForce RTX 4070 GPU with 12 GB of memory to accelerate training. All relevant models are implemented based on the PyTorch 2.0 platform. To ensure the fairness of the experiments, all models are trained, validated, and tested in the same experimental environment. AdamW [44] is utilized as the optimizer. Specifically, the initial learning rates are set to 0.0002 and 0.0003 for the Bijie and Luding datasets, respectively. To dynamically adjust the learning rate throughout the training process, a cosine annealing strategy is employed. The batch size is set to 4, and all models are trained for 60 epochs. For the weighted cross-entropy loss function, the parameters for the background and foreground weights are assigned values of 0.5 and 1.0, respectively. The α in the joint loss function is set to 0.8.

3. Results and Analysis

3.1. Results and Analysis of the Ablation Experiment

To further elaborate on the significance of the dual-branch structure introduced by DBSANet, as well as the FFM and SGAM, this section presents the results of the ablation experiments conducted on the Bijie and Luding datasets, aiming to quantitatively assess the influence of each component on the landslide detection accuracy. The baseline for the ablation experiments is the DBSANet model with the FMM and SGAM removed. In the encoder part, features from the CNN and Transformer branches are combined through pixel-wise addition operations.

3.1.1. Significance of the Dual Branch

To validate the effectiveness of the dual-branch encoder structure adopted by the DBSANet model's encoder, we performed a comparative analysis between DBSANet and variants using either ResNet-18 or Swin-T as the encoder. The results of the experiment are shown in Table 1. Specifically, on the Bijie dataset, the introduction of the dual-branch encoder structure resulted in an average increase of 3.04% in the IoU, a 0.39% average improvement in the accuracy, and a 1.97% average rise in the F1_score. On the Luding landslide dataset, comparable performance enhancements were observed, with the IoU, accuracy, and F1_score improving on average by 2.79%, 1.11%, and 1.85%, respectively.

These results indicate that employing the dual-branch encoder structure with ResNet-18 and Swin-T significantly enhances the accuracy of landslide detection.

Table 1. Comparisons of the two-branch encoder ablation experiment.

Dataset	Model	IoU	Accuracy	F1_Score
Bijie	ResNet-18	74.08%	97.21%	85.11%
	Swin-T	72.67%	97.01%	84.17%
	DBSANet	77.12%	97.60%	87.08%
Luding	ResNet-18	72.44%	91.86%	84.02%
	Swin-T	73.92%	92.42%	85.00%
	DBSANet	75.23%	92.97%	85.87%

To facilitate a more intuitive analysis of the performance of various encoders in landslide detection, we visualized their segmentation results using the statistical values derived from the confusion matrices (TP, FP, and FN) (as shown in Figure 7). As shown in the figure, when only the ResNet-18 encoder is used, its strong local feature extraction capability can capture subtle changes in the image. However, this focused localization occasionally results in misclassifying extraneous backgrounds, such as bare ground, as landslide areas. Moreover, its deficiency in long-range dependencies occasionally results in misidentifying landslides as non-landslide areas (see the red boxes in the ResNet-18 segmentation results in Figure 7). On the contrary, when Swin-T is employed as the encoder, its Transformer-based global context modeling capability is able to capture extensive dependencies between landslides and backgrounds. However, accurately localizing landslide boundaries remains challenging, resulting in a notable increase in FP within the segmentation results (see the red box for the Swin-T segmentation result in Figure 7). When we adopt the dual-branch encoder that combines ResNet-18 and Swin-T, significant alterations in the results of landslide segmentation are observed. This integration harnesses the CNN's capability to capture intricate local details, facilitating the precise identification of minute features within the landslide zone. Additionally, it employs Transformer to establish long-range dependencies, effectively delineating boundaries between landslides and the surrounding terrain, thereby mitigating interference from similar objects [see Figure 7 for DBSANet segmentation results]. Therefore, this further proves the significance of each branch of DBSANet.

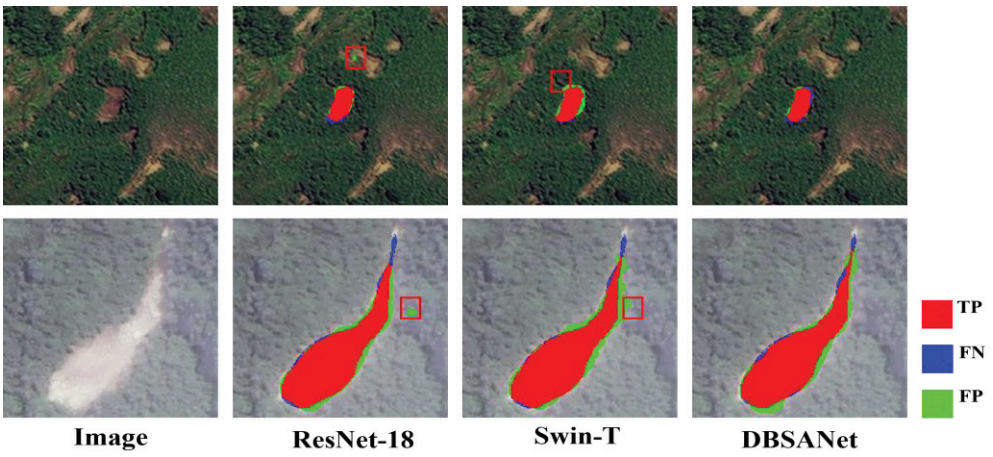


Figure 7. Visualization of confusion matrix statistics for the Bijie and Luding datasets.

3.1.2. Significance of the FFM

These experiments are designed to thoroughly evaluate the impact of the FFM on the accuracy of landslide detection. The experimental results are detailed in Table 2. Using the results from the Bijie dataset as an example, the introduction of the FFM into the baseline model led to improvements in the IoU by 2.72%, accuracy by 0.32%, and F1_score by 1.79%. This enhancement demonstrates the FFM's pivotal role in facilitating effective feature fusion and enhancing the model landslide detection accuracy. Conversely, upon removing the FFM from DBSANet, we observed a contrasting performance trend: the IoU decreased by 2.92%, the accuracy decreased by 0.30%, and the F1_score decreased by 1.89%. These results not only quantitatively demonstrate the beneficial impact of the FFM on the performance of the DBSANet model but also underscore the critical role of effective feature fusion methods in improving the model's overall efficacy.

Table 2. Ablation experiment results of FFM.

Dataset	Model	IoU	Accuracy	F1_Score
Bijie	Baseline	72.91%	97.12%	84.34%
	Baseline + FFM	75.63%	97.44%	86.13%
	DBSANet (no FFM)	74.20%	97.30%	85.19%
	DBSANet	77.12%	97.60%	87.08%
Luding	Baseline	71.74%	91.65%	83.55%
	Baseline + FFM	72.95%	91.90%	84.36%
	DBSANet (no FFM)	73.73%	92.54%	84.88%
	DBSANet	75.23%	92.97%	85.87%

To elucidate the specific role of the FFM, we conducted a visualization of the extracted feature maps, as shown in Figure 8. Simply adding the features extracted by both models could result in redundancy or conflicts due to feature inconsistencies, potentially impacting the overall accuracy of landslide detection. As is clearly visible in the figure, the removal of the FFM in DBSANet results in notable changes in its visualization effects. Specifically, the feature values in the image display a dispersed and irregular distribution, lacking distinct concentration, which hampers the model's ability to effectively target the critical landslide-related regions. In contrast, results from DBSANet reveal pronounced responses in landslide-related areas, with the background predominantly composed of blue regions, suggesting enhanced model focus on these critical areas. In summary, the application of the FFM in DBSANet enhances feature fusion, thereby enabling the more precise capture of landslide features by the model.

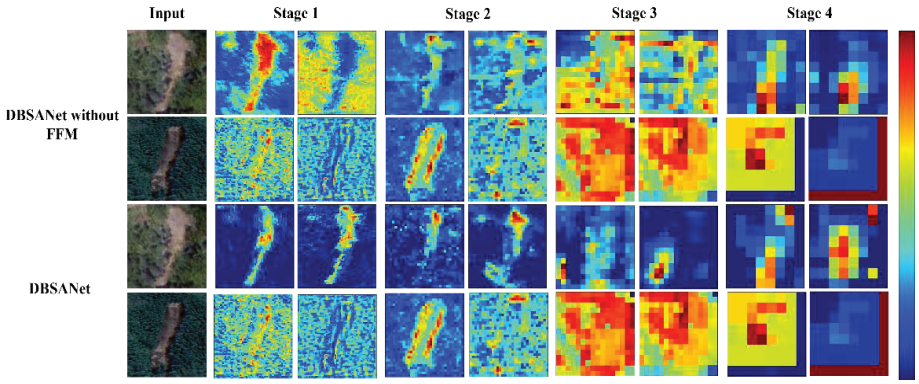


Figure 8. An example of the feature map COLORMAP_JET pseudocolor visualization in the Luding dataset, where Stage 1, Stage 2, Stage 3, and Stage 4 correspond to the four stages of feature extraction for the four backbone encoders (any two output channels).

3.1.3. Significance of the SGAM

The experimental results are summarized in Table 3. The results from the Bijie dataset reveal a 1.29% improvement in the IoU, a 0.18% increase in the accuracy, and a 0.85% enhancement in the F1_score following the integration of the SGAM into the baseline model. Conversely, upon removing the SGAM from DBSANet, there was a decline in performance metrics: the IoU decreased by 1.49%, the accuracy dropped by 0.16%, and the F1_score decreased by 0.95%. This consistent trend in performance indicators is similarly observed in the Luding dataset. These comparative experiments collectively underscore the efficacy of the SGAM in bridging the semantic gap between the encoder and decoder, highlighting its beneficial impact within DBSANet.

Table 3. Ablation experiment results of SGAM.

Dataset	Model	IoU	Accuracy	F1_Score
Bijie	Baseline	72.91%	97.12%	84.34%
	Baseline + SGAM	74.20%	97.30%	85.19%
	DBSANet (no SGAM)	75.63%	97.44%	86.13%
	DBSANet	77.12%	97.60%	87.08%
Luding	Baseline	71.74%	91.65%	83.55%
	Baseline + SGAM	73.73%	92.54%	84.88%
	DBSANet (no SGAM)	72.95%	91.90%	84.36%
	DBSANet	75.23%	92.97%	85.87%

To provide a clearer illustration of the encoder–decoder semantic gap, we compare and present the visualization results of a selected output channel from Encoder Stage 1 and its corresponding counterpart in the decoder in Figure 9. As depicted in panels (e) and (f) of Figure 9, the encoder's shallow feature maps are rich in intricate low-level image features such as shapes and textures. Conversely, the decoder contains a multitude of coarse-grained high-level image features that can directly distinguish between categories, resulting in more abstract image visualization outcomes. When attempting to simply add these two types of features together (as shown in Figure 9g), it became evident that the activation values in certain non-landslide regions were unusually elevated. Conversely, following processing by the SGAM, the visualization of the output channel predominantly exhibited blue hues in the background areas. Concurrently, we utilized gradient-weighted class activation mapping (Grad-CAM) [45] as an analytical tool, aimed at providing a more intuitive depiction of SGAM's role in bridging the semantic gap. Figure 9 illustrates that Grad-CAM vividly depicts the model's attention distribution across different image regions during inference. Specifically, in Figure 9c, we observe that without the SGAM in DBSANet, the model focuses on the landslide area but also erroneously assigns higher activation values to background roads that share similar color tones with the landslide area (highlighted in the red box). In contrast, as illustrated in Figure 9d, with the incorporation of the SGAM, the model exhibits enhanced precision in its focus. It not only intensifies its attention on the landslide region but also effectively mitigates background noise, leading to clearer and more precise segmentation results. This improvement underscores the SGAM's effectiveness in bridging the semantic gap between the encoder and decoder.

3.2. Results and Analysis of Landslide Extraction

3.2.1. Experiment 1—Bijie Dataset

As shown in Table 4, DBSANet has attained a notable performance on the Bijie dataset, achieving IoU, accuracy, and F1-score values of 77.12%, 97.60%, and 87.08%, respectively. To conduct a thorough evaluation of DBSANet's performance, we employed Grad-CAM alongside statistical metrics from the confusion matrix (FN, FP, and TP) to visualize the seg-

mentation outcomes, as illustrated in Figure 10. It can be observed that when the landslide boundary is clearly defined and significantly different from the surrounding terrain texture (I and II in Figure 10), DBSANet is able to accurately capture and segment the landslide area. Meanwhile, in the face of complex backgrounds with similar colors and textures and without significant interference from roads or buildings, DBSANet remains focused on the landslide itself, effectively avoiding misjudgments and omissions. However, in scenarios where the landslide area includes additional topographical features, as illustrated in Figure 10III, such as the presence of new vegetation, accurately delineating the boundaries of these mixed areas proves challenging. Despite Grad-CAM indicating the model's focus on the core landslide region, occasional misclassifications occur where adjacent vegetated zones are mistakenly identified as landslides, thereby increasing the TP value.

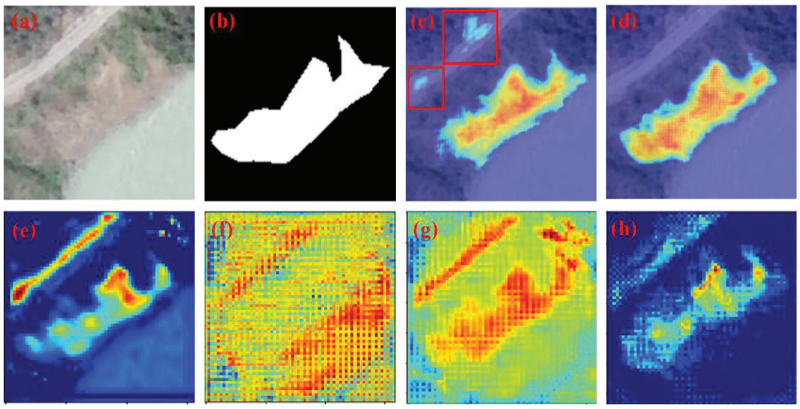


Figure 9. Bijie dataset image Grad-CAM visualization (**c**,**d**) and COLORMAP_JET pseudocolor visualization (**e**,**h**). (**a**) Image. (**b**) Label. (**c**) Grad-CAM visualization of DBSANet (without SGAM) output layer. (**d**) Grad-CAM visualization of DBSANet output layer. (**e**) DBSANet Encoder Stage 1 output channel visualization. (**f**) Decoder output channel visualization corresponding to Stage 1 in DBSANet. (**g**) DBSANet (without SGAM) output channel visualization. (**h**) DBSANet output channel visualization.

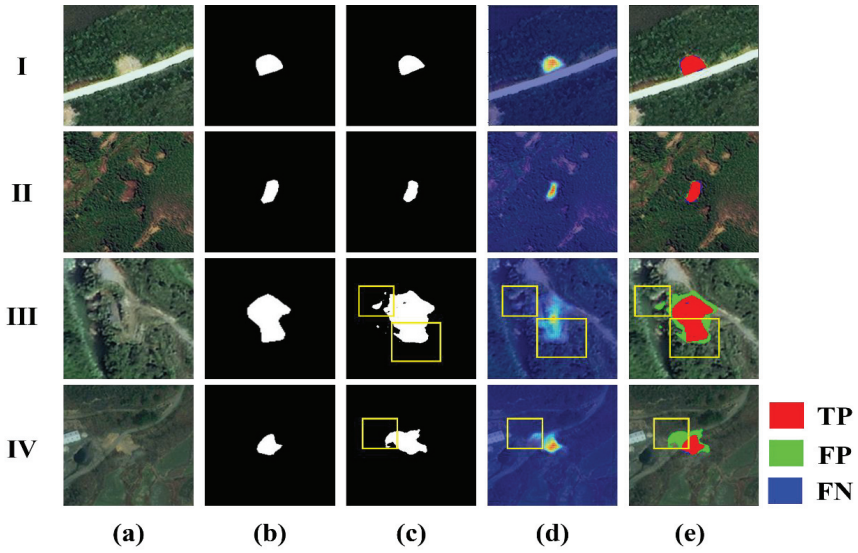


Figure 10. Test results of DBSANet on Bijie test set (**I–IV** are four example images in the test set). (**a**) Image. (**b**) Label. (**c**) DBSANet segmentation result. (**d**) Grad-CAM visualization of DBSANet's output layer. (**e**) Visualization of confusion matrix statistics. TP denotes the number of pixels accurately identified as landslides; FP denotes the number of pixels where the background is incorrectly recognized as a landslide; and FN denotes the number of pixels where the landslide is incorrectly recognized as background.

Table 4. Accuracy of DBSANet in Bijie and Luding datasets.

Dataset	IoU	Accuracy	F1_Score
Bijie	77.12%	97.60%	87.08%
Luding	75.23%	92.97%	85.87%

3.2.2. Experiment 2—Luding Dataset

As shown in Table 4, DBSANet achieved an IoU, accuracy, and F1_score of 75.23%, 92.97%, and 85.87%, respectively, on the Luding dataset. We also visualized its results, as shown in Figure 11. The lower spatial resolution of this dataset compared to the Bijie dataset poses more challenges for accurately locating landslide boundaries. As depicted in Figure 11, I–II, when the surface texture around the landslide is significantly different from the landslide itself, the main body of the landslide can be correctly segmented, with fewer misjudgments in the boundary area (as indicated by the yellow box). However, in scenarios where new vegetation emerges around the landslide (III and IV in Figure 11), the topographical intricacies of the landslide surface heighten, which directly increases the difficulty of accurately identifying the landslide boundary. Therefore, when dealing with images with low spatial resolution, DBSANet still faces challenges in accurately locating landslide boundaries solely based on optical images as input.

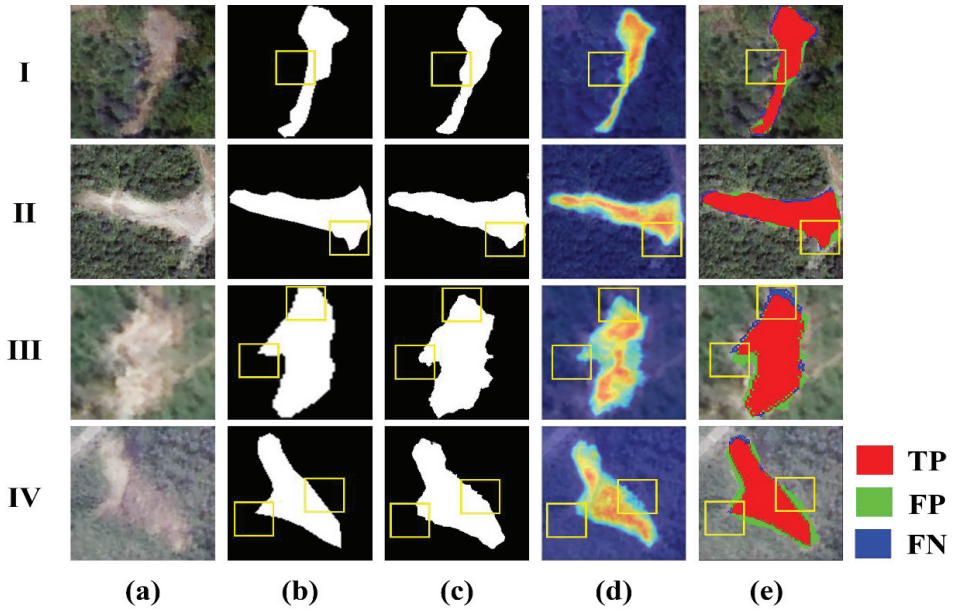


Figure 11. Test results of DBSANet on Luding test set (**I–IV** are four example images in the test set). (**a**) Image. (**b**) Label. (**c**) DBSANet segmentation result. (**d**) Grad-CAM visualization of DBSANet's output layer. (**e**) Visualization of confusion matrix statistics. TP denotes the number of pixels accurately identified as landslides; FP denotes the number of pixels where the background is incorrectly recognized as a landslide; and FN denotes the number of pixels where the landslide is incorrectly recognized as background.

4. Discussion

To comprehensively assess the performance of DBSANet, this study selects seven classic semantic segmentation models for comparative analysis, organizing them into three distinct categories according to their architectural features. The initial category comprises models that are exclusively based on CNNs, exemplified by UNet, Deeplabv3+, and ResUNet. The second category is fully based on the Transformer architecture, represented by SwinUNet. The third category is a hybrid of a CNN and Transformer, including TransUNet,

TransFuse, and UNetFormer. Through these comparative experiments, our objective is to conclusively showcase the superior efficacy of DBSANet in landslide detection tasks.

4.1. Comparison with Other CNNs

Tables 5 and 6 present the experimental results for various models evaluated on the Bijie dataset and the Luding dataset. The data clearly indicate that DBSANet outperforms all other models across the three performance metrics. Specifically, compared to the other seven models for comparison, DBSANet exhibits average enhancements of 4.91% in the IoU, 0.57% in the accuracy, and 3.25% in the F1_score. Likewise, on the Luding dataset, DBSANet displays exceptional performance, demonstrating average improvements of 2.96%, 1.08%, and 1.97% in the IoU, accuracy, and F1_score, respectively. It has achieved superior performance on both the Bijie and Luding datasets, further suggesting that our model can more precisely align the overlapping regions between the actual and predicted landslide areas during detection. This not only reduces the false alarm rate, defined as the incorrect identification of non-landslide areas as landslide areas, but also lowers the missed detection rate, which refers to the failure to identify actual landslide regions. To visually compare the landslide detection performance of DBSANet with other models, we visualized the prediction outcomes for each model, as depicted in Figures 12 and 13. The figure clearly illustrates that true positives (TPs) dominate the area in the DBSANet visualization results, while false positives (FPs) and false negatives (FNs) occupy comparatively smaller areas, thereby effectively demonstrating the network's efficacy in the landslide detection task.

Table 5. Comparisons of accuracy among various models on the Bijie dataset.

Model	IoU	Accuracy	F1_Score
UNet	73.93%	97.20%	85.01%
Deeplabv3+	71.74%	97.07%	83.55%
ResUNet	74.74%	97.36%	85.55%
SwinUNet	66.31%	96.22%	79.74%
TransUNet	72.09%	97.08%	83.78%
TransFuse	72.89%	97.06%	84.32%
UNetFormer	73.75%	97.19%	84.89%
DBSANet	77.12%	97.60%	87.08%

Table 6. Comparisons of accuracy among various models on the Luding dataset.

Model	IoU	Accuracy	F1_Score
UNet	73.04%	91.96%	84.42%
Deeplabv3+	70.75%	91.36%	82.87%
ResUNet	72.20%	92.03%	83.86%
SwinUNet	72.57%	92.06%	84.10%
TransUNet	72.04%	91.85%	83.75%
TransFuse	72.50%	91.93%	84.06%
UNetFormer	72.76%	92.02%	84.23%
DBSANet	75.23%	92.97%	85.87%

In the experiments conducted on the Bijie dataset, ResUNet exhibited a superior performance among the CNN-based models when compared to UNet and Deeplabv3+, indicating that the adoption of the ResNet structure can achieve a better feature representation on the Bijie dataset. Compared to traditional CNN-based methods, SwinUNet introduces a self-attention mechanism for feature extraction. Although it adopts a similar network structure to UNet, its relatively simple network architecture somewhat limits its performance in landslide detection tasks. As a consequence, its overall performance falls

short when compared to other benchmark models. The segmentation results depicted in Figure 12 reveal a degree of boundary blurring in the landslide region when using SwinUNet. This observation suggests that SwinUNet's reliance on a self-attention mechanism for global information interaction, in contrast to CNNs, might compromise the precision of local detail processing. Conversely, TransUNet, TransFuse, and UNetFormer adeptly amalgamate the CNN's and Transformer's strengths, albeit through distinct integration strategies. TransUNet employs a hybrid encoder strategy. Initially, CNNs are utilized for feature extraction, followed by partitioning the resulting feature maps into patches and transforming them into one-dimensional vectors. These vectors are then integrated into the Transformer framework to enhance global comprehension. In contrast, TransFuse concurrently processes the CNN's and Transformer's features, employing novel fusion techniques to effectively amalgamate features derived from parallel branches. UNetFormer follows a structure akin to UNet but enhances the fusion of global and local information by substituting the decoder with a Transformer. These hybrid models demonstrated significant improvements in IoU metrics—5.78%, 6.58%, and 7.44%, respectively—compared to SwinUNet's pure Transformer architecture, underscoring the efficacy of hybrid model architectures in landslide detection tasks. In comparison to the previously mentioned CNN-based models, pure Transformer architecture models, and hybrid models, DBSANet, as proposed in this study, demonstrates notable performance advantages in the context of landslide detection tasks. More specifically, when compared to the top-performing models within each respective category (ResUNet, SwinUNet, and UNetFormer), DBSANet exhibits improvements of 2.38%, 10.81%, and 3.37% in the IoU metric performance, thus unequivocally showcasing DBSANet's superiority in the realm of landslide detection tasks.

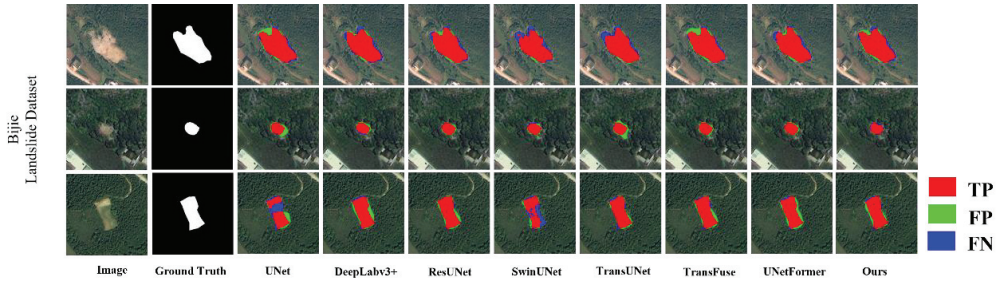


Figure 12. Visualization of confusion matrix statistics for the Bijie dataset.

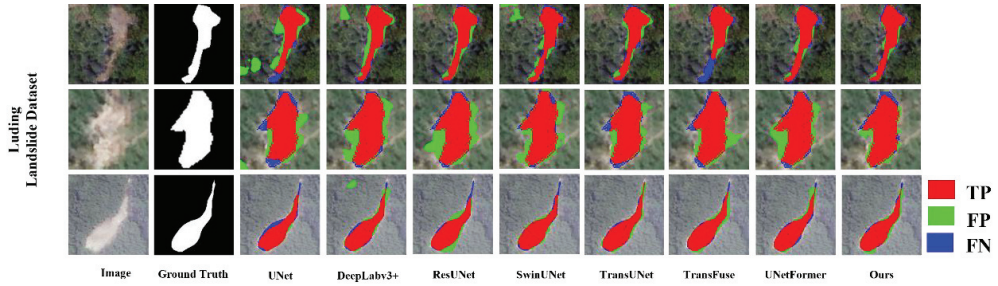


Figure 13. Visualization of confusion matrix statistics for the Luding dataset.

In the Luding dataset experiments, while comparison models like UNet stand out among CNN-based models and UNetFormer achieves the top performance among the CNN and Transformer hybrid models, their evaluation metrics still do not surpass those of DBSANet proposed in this study. Particularly noteworthy is that, compared to the top-performing model in each category, the DBSANet proposed in this study has achieved improvements in the IoU values of 2.19%, 2.66%, and 2.47%, respectively. This underscores

the broad applicability of DBSANet across various landslide datasets and its efficacy in landslide detection.

4.2. Comparison of Model Efficiency

To facilitate a clear comparison of computational efficiency among different models, this study quantifies the computational efficiency based on the model parameters, average training time for each epoch, and inference speed per image, with the experimental results detailed in Table 7. As illustrated in the table, DBSANet does not lead to excessive memory consumption; however, it is worth noting that the dual-branch structure adopted by the model also enhances the training time of the model to some extent. Compared to models based on CNNs or pure Transformer structures, DBSANet makes a trade-off between efficiency and accuracy, but leans more towards the latter. Furthermore, in contrast to the hybrid model, DBSANet attains superior results in landslide detection using fewer parameters.

Table 7. Comparisons of parameters and inference speed among various models.

Model	Parameters (M)	Average Training Time (s/Epoch)	Inference Speed (ms/Image)
UNet	17.23	20.73	4.85
Deeplabv3+	59.34	20.76	6.10
ResUNet	48.55	14.33	5.28
Swin_UNet	27.19	15.06	4.18
TransUNet	51.07	22.41	5.17
TransFuse	98.49	24.06	7.87
UNetFormer	21.83	15.58	3.02
Ours	39.98	25.56	6.69

4.3. Application of DBSANet in Other Scenarios

In order to further validate the generalization of the DBSANet model, we identified the Rizhaigou watershed in Jiuzhaigou County as the experimental site. The region's abundant rainfall, in conjunction with its intricate geological formations, frequently precipitates the occurrence of landslides alongside the rivers. In the preliminary phase of the experiment, we successfully obtained landslide samples from this area by combining GF-2 satellite imagery with Google Earth imagery through visual interpretation. Subsequently, building upon the training results from the Luding dataset, we fine-tuned the model and performed an extensive evaluation on GF-2 imagery obtained from the Rizhaigou watershed. The experimental results are shown in Figure 14.

As illustrated in the figure, the DBSANet model effectively identified the majority of landslide areas within the experimental region. Specifically, the model's performance metrics were as follows: the IoU was 66.52%, the accuracy was 95.13%, and the F1-score was 79.90%. However, it is important to note that while these metrics reflect a certain performance level, they have not yet reached optimal levels. This is primarily due to the fact that applying a trained model to an unexplored region presents a considerable challenge. Most prior studies have conducted training and validation within the same region, thus avoiding challenges such as geological structure disparities and variations in image features. In contrast, the images in the Luding dataset were captured shortly after an earthquake, displaying notable differences from the test images of the Rizhaigou watershed (acquired in May 2023). Furthermore, as observed in the prediction results presented in Figure 14, the DBSANet model erroneously classified non-landslide areas as landslides in some instances (indicated by red boxes). This misidentification may arise from two primary factors: first, the manual labeling of landslide tags inherently involves

subjectivity, and labeling errors may lead the model to acquire erroneous information; second, the majority of landslides in the Rizhaigou watershed are aged, and the vegetation on the slopes has developed over an extended period, causing the landslide boundaries to become blurred and difficult to define accurately. This alteration in vegetation cover undoubtedly exacerbates the challenges associated with landslide detection. In summary, while the DBSANet model has shown the capacity to accurately identify landslide positions in other application contexts, the experimental results from the Rizhaigou watershed show that the model still requires improvement in determining the specific scope of landslides.

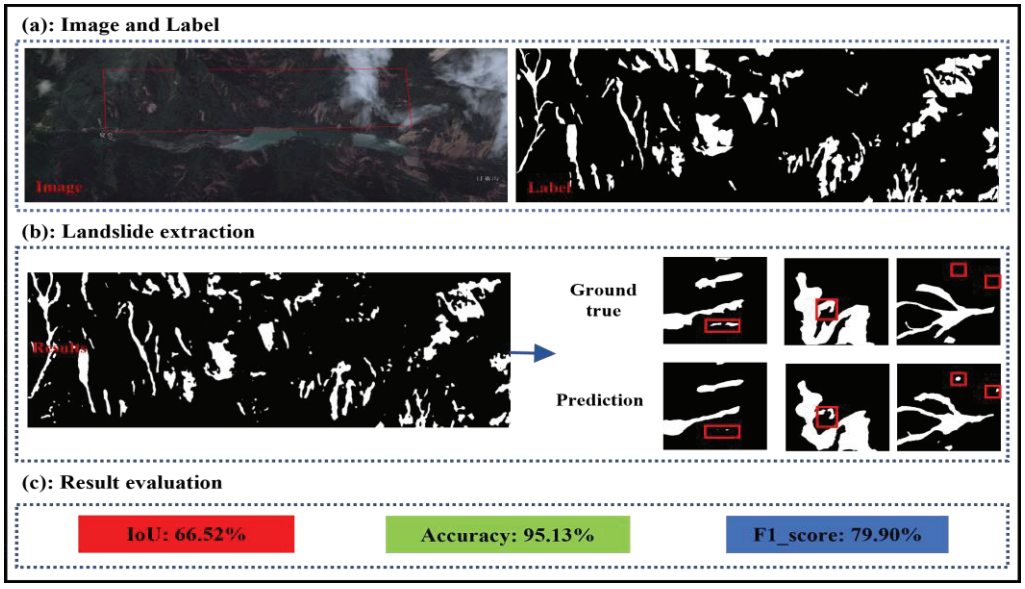


Figure 14. Application of DBSANet in other scenarios.

4.4. Limitations and Future Work

The DBSANet model employs a dual-branch parallel structure consisting of a ResNet and Swin Transformer as its encoder. The Swin Transformer branch effectively captures long-range dependencies across various regions of the image, whereas the ResNet branch extracts local feature information, thereby facilitating the model's acquisition of more comprehensive feature representations. From Tables 5 and 6, it is evident that the model demonstrates a substantial enhancement in landslide detection performance. However, the analysis of Figures 10 and 11 reveals the difficulties associated with accurately delineating the extent of landslides, particularly when they are incomplete, such as during the emergence of nascent vegetation. This limitation can be attributed to the restricted sample size of the landslide dataset employed in this study, which impedes the model's ability to represent all possible types of landslides. Consequently, the emergence of new vegetation on the landslide surface presents a significant challenge for the model, hindering its ability to precisely delineate the landslide boundaries.

In future research, we will gather landslide data across various scenarios to enhance the diversity of features available for model training. Furthermore, to improve the generalizability and adaptability of the model across various application scenarios, we will integrate a diverse range of multi-source remote sensing data—including, but not limited to, digital elevation models, and slope information—into the dataset construction, which is anticipated to further enhance the model's performance in real-world contexts, given the substantial influence of these factors on landslide occurrences.

5. Conclusions

In this study, we propose a semantic segmentation model named DBSANet, which is based on a U-shaped encoder–decoder architecture for landslide detection in remote sensing images. Inspired from the Transformer's global modeling via self-attention and the CNN's adept feature extraction capabilities, the DBSANet encoder integrates a dual-branch parallel structure of a Transformer and CNN. This design facilitates the concurrent acquisition of both comprehensive contextual information and low-level spatial details of images. Given the inherent structural disparities in feature representations between the CNN and Transformer during extraction, we propose an FFM designed to merge features from both branches, thereby effectively integrating their complementary attributes. Additionally, in order to bridge the semantic gap between the encoder and decoder and thereby enhance the overall model performance, we design an SGAM integrated within the skip connection segment. This module aims to heighten the model's emphasis on pivotal areas by aligning the decoder feature maps with those of the encoder, effectively bridging the semantic gap between them. To thoroughly and systematically assess the criticality of each module within the model, we carefully designed and executed a series of ablation experiments. The final results of the experiments unequivocally demonstrate that these modules play indispensable roles. To further verify the efficacy and generalizability of DBSANet, experiments were performed on the Bijie and Luding datasets. The experimental results demonstrated that DBSANet markedly exceeded the performance of the other comparative models in terms of the semantic segmentation accuracy. This further demonstrates that the meticulously designed DBSANet in this study effectively integrates the advantages of CNNs and Transformers, thereby enhancing the accuracy of landslide detection, and also provides substantial technical support for the deployment of the hybrid model in landslide detection applications.

Although the remarkable performance exhibited by DBSANet in a variety of experiments is impressive, two limitations persist. First, the incorporation of a dual-branch encoder inherently leads to an increase in the number of parameters, thereby extending the training duration. Second, in scenarios where the surface of landslide bodies is obscured by emerging vegetation, DBSANet encounters difficulties in precisely delineating the landslide boundaries. To address these limitations, future research will concentrate on enhancing efficiency while preserving high accuracy, and applying this equilibrium to landslide detection tasks. Furthermore, we plan to explore ways to integrate multi-source data into the landslide detection process, ensuring that the model can effectively learn and extract crucial features in diverse, complex scenarios, thereby enhancing the recognition of landslide boundaries and improving the model's generalization performance.

Author Contributions: All the authors made significant contributions to this work. Conceptualization, W.Z. and Y.L.; methodology, W.Z. and Y.L.; software and experiments, Y.L.; validation, J.W., R.Z. and Y.Z.; writing—original draft preparation, Y.L.; writing—review and editing, W.Z. and X.X.; supervision, Y.Z.; funding acquisition, W.Z. and X.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 (2023YFC3008402), Natural Science Foundation of China (NSFC) (Nos.: 42474051, 42074040), Natural Science Basic Research Plan in Shaanxi Province of China (Nos.: 2023-JC-JQ-24, 2022JQ-454), the innovation team of Shaanxi Provincial Tri-Qin Scholars with Geoscience Big Data and Geohazard Prevention (2022), Fundamental Research Funds for the Central Universities of CHD (Nos.: 300102263401, 300102264502, 300102264915).

Data Availability Statement: The Bijie dataset used in this study can be obtained from the Bijie Landslide Dataset (whu.edu.cn), accessed on 30 March 2024. The satellite imagery data involved in the study can be acquired from the China Center for Resources Satellite Data and Application at https://www.cresda.com/zgzywxyyzx/index.html, accessed on 1 November 2024.

Acknowledgments: The authors would like to express their sincere gratitude to the reviewers and editors for their constructive and high-quality revision suggestions on this manuscript.

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

References

1. Zhu, W.; Yang, L.Y.; Cheng, Y.Q.; Liu, X.Y.; Zhang, R.X. Active thickness estimation and failure simulation of translational landslide using multi-orbit InSAR observations: A case study of the Xiongba landslide. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *129*, 16. [CrossRef]
2. Pardeshi, S.D.; Autade, S.E.; Pardeshi, S.S. Landslide hazard assessment: Recent trends and techniques. *SpringerPlus* **2013**, *2*, 523. [CrossRef] [PubMed]
3. Zhang, J.M.; Zhu, W.; Cheng, Y.Q.; Li, Z.H. Landslide Detection in the Linzhi-Ya'an Section along the Sichuan-Tibet Railway Based on InSAR and Hot Spot Analysis Methods. *Remote Sens.* **2021**, *13*, 3566. [CrossRef]
4. Wang, C.S.; Chang, L.; Wang, X.S.; Zhang, B.C.; Stein, A. Interferometric Synthetic Aperture Radar Statistical Inference in Deformation Measurement and Geophysical Inversion: A review. *IEEE Geosci. Remote Sens. Mag.* **2024**, *12*, 8–35. [CrossRef]
5. Zhang, R.-X.; Zhu, W.; Li, Z.-H.; Zhang, B.-C.; Chen, B. Re-Net: Multibranch Network With Structural Reparameterization for Landslide Detection in Optical Imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 2828–2837. [CrossRef]
6. Ansari, M.D.; Ghrera, S.; Tyagi, V. Pixel-Based Image Forgery Detection: A Review. *IETE J. Educ.* **2014**, *55*, 40–46. [CrossRef]
7. Ghorbanzadeh, O.; Gholamnia, K.; Ghamisi, P. The application of ResU-net and OBIA for landslide detection from multi-temporal sentinel-2 images. *Big Earth Data* **2023**, *7*, 961–985. [CrossRef]
8. Shi, W.Z.; Zhang, M.; Ke, H.F.; Fang, X.; Zhan, Z.; Chen, S.X. Landslide Recognition by Deep Convolutional Neural Network and Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 4654–4672. [CrossRef]
9. Cheng, L.B.; Li, J.; Duan, P.; Wang, M.G. A small attentional YOLO model for landslide detection from satellite remote sensing images. *Landslides* **2021**, *18*, 2751–2765. [CrossRef]
10. Ullo, S.L.; Mohan, A.; Sebastianelli, A.; Ahamed, S.E.; Kumar, B.; Dwivedi, R.; Sinha, G. 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]
11. Noh, H.; Hong, S.; Han, B. Learning Deconvolution Network for Semantic Segmentation. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 11–18 December 2015; pp. 1520–1528.
12. Yi, Y.; Zhang, W. A New Deep-Learning-Based Approach for Earthquake-Triggered Landslide Detection From Single-Temporal RapidEye Satellite Imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 6166–6176. [CrossRef]
13. Prakash, N.; Manconi, A.; Loew, S. Mapping Landslides on EO Data: Performance of Deep Learning Models vs. Traditional Machine Learning Models. *Remote Sens.* **2020**, *12*, 346. [CrossRef]
14. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
15. He, K.M.; Zhang, X.Y.; Ren, S.Q.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 27–30 June 2016; pp. 770–778.
16. Meena, S.R.; Soares, L.P.; Grohmann, C.H.; van Westen, C.; Bhuyan, K.; Singh, R.P.; Floris, M.; Catani, F. Landslide detection in the Himalayas using machine learning algorithms and U-Net. *Landslides* **2022**, *19*, 1209–1229. [CrossRef]
17. Li, Z.; Guo, Y. Semantic segmentation of landslide images in Nyingchi region based on PSPNet network. In Proceedings of the 2020 7th International Conference on Information Science and Control Engineering (ICISCE), Changsha, China, 8–20 December 2020; pp. 1269–1273.
18. Qi, W.W.; Wei, M.F.; Yang, W.T.; Xu, C.; Ma, C. Automatic Mapping of Landslides by the ResU-Net. *Remote Sens.* **2020**, *12*, 2487. [CrossRef]
19. 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 15th European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; Proceedings, Part VII; pp. 833–851.
20. Xia, W.; Chen, J.; Liu, J.B.; Ma, C.H.; Liu, W. Landslide Extraction from High-Resolution Remote Sensing Imagery Using Fully Convolutional Spectral-Topographic Fusion Network. *Remote Sens.* **2021**, *13*, 5116. [CrossRef]
21. Chen, H.S.; He, Y.; Zhang, L.F.; Yao, S.; Yang, W.; Fang, Y.M.; Liu, Y.X.; Gao, B.H. A landslide extraction method of channel attention mechanism U-Net network based on Sentinel-2A remote sensing images. *Int. J. Digit. Earth* **2023**, *16*, 552–577. [CrossRef]

22. Zhou, Z.W.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J.M. UNet plus plus: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation. *IEEE Trans. Med. Imaging* **2020**, *39*, 1856–1867. [CrossRef]
23. Pang, Y.W.; Li, Y.Z.; Shen, J.B.; Shao, L. Towards Bridging Semantic Gap to Improve Semantic Segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 4229–4238.
24. 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 31st Annual Conference on Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 4–9 December 2017.
25. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. *Swin-Unet: Unet-Like Pure Transformer for Medical Image Segmentation*; Springer: Cham, Switzerland, 2021.
26. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. *arXiv* **2021**, arXiv:2102.04306.
27. Zhang, Y.D.; Liu, H.Y.; Hu, Q. TransFuse: Fusing Transformers and CNNs for Medical Image Segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), Strasbourg, France, 27 September–1 October 2021; pp. 14–24.
28. Wang, L.B.; Li, R.; Zhang, C.; Fang, S.H.; Duan, C.X.; Meng, X.L.; Atkinson, P.M. UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. ISPRS-J. Photogramm. *Remote Sens.* **2022**, *190*, 196–214. [CrossRef]
29. Lv, P.Y.; Ma, L.S.; Li, Q.M.; 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]
30. Tang, X.C.; Tu, Z.H.; Wang, Y.; Liu, M.Z.; Li, D.F.; Fan, X.M. Automatic Detection of Coseismic Landslides Using a New Transformer Method. *Remote Sens.* **2022**, *14*, 2884. [CrossRef]
31. Azad, R.; Heidari, M.; Wu, Y.L.; Merhof, D. Contextual Attention Network: Transformer Meets U-Net. In Proceedings of the 13th International Workshop on Machine Learning in Medical Imaging (MLMI), Singapore, 18 September 2022; pp. 377–386.
32. Xiang, X.Y.; Gong, W.P.; Li, S.L.; Chen, J.; Ren, T.H. TCNet: Multiscale Fusion of Transformer and CNN for Semantic Segmentation of Remote Sensing Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 3123–3136. [CrossRef]
33. Li, J.; Zhang, J.; Fu, Y.Y. CTHNet: A CNN-Transformer Hybrid Network for Landslide Identification in Loess Plateau Regions Using High-Resolution Remote Sensing Images. *Sensors* **2025**, *25*, 273. [CrossRef] [PubMed]
34. Yang, Z.Q.; Xu, C.; Li, L. Landslide Detection Based on ResU-Net with Transformer and CBAM Embedded: Two Examples with Geologically Different Environments. *Remote Sens.* **2022**, *14*, 2885. [CrossRef]
35. Ji, S.P.; Yu, D.W.; Shen, C.Y.; Li, W.L.; Xu, Q. Landslide detection from an open satellite imagery and digital elevation model dataset using attention boosted convolutional neural networks. *Landslides* **2020**, *17*, 1337–1352. [CrossRef]
36. Liu, Z.; Lin, Y.T.; Cao, Y.; Hu, H.; Wei, Y.X.; Zhang, Z.; Lin, S.; Guo, B.N. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In Proceedings of the 18th IEEE/CVF International Conference on Computer Vision (ICCV), 11–17 October 2021; pp. 9992–10002.
37. Jiang, L.; Zhang, C.; Huang, M.; Liu, C.; Shi, J.; Loy, C.C. TSIT: A Simple and Versatile Framework for Image-to-Image Translation. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; pp. 206–222.
38. Wang, X.T.; Yu, K.; Dong, C.; Loy, C.C. Recovering Realistic Texture in Image Super-resolution by Deep Spatial Feature Transform. In Proceedings of the 31st IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 606–615.
39. Lee, H.; Kim, H.E.; Nam, H. SRM: A Style-based Recalibration Module for Convolutional Neural Networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1854–1862.
40. Howard, A.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv* **2017**, arXiv:1704.04861. [CrossRef]
41. Keyport, R.N.; Oommen, T.; Martha, T.R.; Sajinkumar, K.S.; Gierke, J.S. A comparative analysis of pixel- and object-based detection of landslides from very high-resolution images. *Int. J. Appl. Earth Obs. Geoinf.* **2018**, *64*, 1–11. [CrossRef]
42. Phan, T.H.; Yamamoto, K. Resolving Class Imbalance in Object Detection with Weighted Cross Entropy Losses. *arXiv* **2020**, arXiv:2006.01413.
43. Milletari, F.; Navab, N.; Ahmadi, S.A. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In Proceedings of the 4th IEEE International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016; pp. 565–571.

44. Kingma, D.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the International Conference on Learning Representations, Banff, AB, Canada, 14–16 April 2014.
45. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. In Proceedings of the 16th IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 618–626.

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

Smoke and Fire-You Only Look Once: A Lightweight Deep Learning Model for Video Smoke and Flame Detection in Natural Scenes

Chenmeng Zhao [1,2,3,4], Like Zhao [5,6], Ka Zhang [1,2,3,4,*], Yinghua Ren [7], Hui Chen [8] and Yehua Sheng [1,2,3,4]

1 Key Laboratory of Virtual Geographic Environment, Ministry of Education, Nanjing Normal University, Nanjing 210023, China; 231312029@njnu.edu.cn (C.Z.); shengyehua@njnu.edu.cn (Y.S.)
2 School of Geography, Nanjing Normal University, Nanjing 210023, China
3 Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing 210023, China
4 State Key Laboratory of Geographical Environment Evolution, Nanjing 210023, China
5 Geological Survey of Jiangsu Province, Nanjing 210018, China; surveyjsgs@163.com
6 Natural Resources Satellite Application Technology Center of Jiangsu Province, Nanjing 210018, China
7 Jiangsu Suzhou Geological Engineering Investigation Institute, Suzhou 215129, China; szgk@szgk.com
8 School of Marine Science and Engineering, Nanjing Normal University, Nanjing 210023, China; huichen@njnu.edu.cn
* Correspondence: 09323@njnu.edu.cn; Tel.: +86-025-13914700581

Abstract: Owing to the demand for smoke and flame detection in natural scenes, this paper proposes a lightweight deep learning model, SF-YOLO (Smoke and Fire-YOLO), for video smoke and flame detection in such environments. Firstly, YOLOv11 is employed as the backbone network, combined with the C3k2 module based on a two-path residual attention mechanism, and a target detection head frame with an embedded attention mechanism. This combination enhances the response of the unobscured regions to compensate for the feature loss in occluded regions, thereby addressing the occlusion problem in dynamic backgrounds. Then, a two-channel loss function (W-SIoU) based on dynamic tuning and intelligent focusing is designed to enhance loss computation in the boundary regions, thus improving the YOLOv11 model's ability to recognize targets with ambiguous boundaries. Finally, the algorithms proposed in this paper are experimentally validated using the self-generated dataset S-Firedata and the public smoke and flame virtual dataset M^4SFWD. These datasets are derived from internet smoke and flame video frame extraction images and open-source smoke and flame dataset images, respectively. The experimental results demonstrate, compared with deep learning models such as YOLOv8, Gold-YOLO, and Faster-RCNN, the SF-YOLO model proposed in this paper is more lightweight and exhibits higher detection accuracy and robustness. The metrics mAP50 and mAP50-95 are improved by 2.5% and 2.4%, respectively, in the self-made dataset S-Firedata, and by 0.7% and 1.4%, respectively, in the publicly available dataset M^4SFWD. The research presented in this paper provides practical methods for the automatic detection of smoke and flame in natural scenes, which can further enhance the effectiveness of fire monitoring systems.

Keywords: natural scenes; smoke and flame detection; video images; attention mechanism; SF-YOLO

1. Introduction

In the context of global climate change, extreme weather events are becoming increasingly prevalent, and the frequency and intensity of fires are escalating. These conflagrations

have resulted in significant casualties and property damage, as well as far-reaching ecological, environmental, and social ramifications. A notable example is the 2023 Bromo fire incident in Indonesia, which was triggered by the use of pyrotechnics by tourists, destroying over 504 hectares of forests and significant ecological and economic damage [1]. Another salient example is the 2023 gas leakage explosion in a civil building in Yinchuan, China, which underscores the critical importance of advanced smoke and flame detection technologies [2]. Consequently, the extraction of smoke and flame information (e.g., location, spreading trend) from diverse surveillance video and remote sensing image types is imperative for effective disaster emergency response and management.

In natural environments, the occurrence of fire events is frequently influenced by a multitude of factors, including drought, wind, and the distribution of combustible materials. This complexity impedes the detection of smoke and flames, particularly under conditions of low contrast or variable lighting. Consequently, conventional smoke and flame detection algorithms, which are predominantly image-based, often demonstrate suboptimal adaptability and detection accuracy under these conditions. Contemporary approaches to smoke and flame detection continue to rely on traditional methods, such as threshold detection, optical sensors, and thermal imaging techniques. Threshold detection involving the setting of specific pixel thresholds to recognize flames, is susceptible to variations in light and background, leading to elevated false alarm rates and proving particularly ineffective in complex environments [3,4]. Optical sensors, which monitor flames based on changes in light, are suitable for scenes with a clear field of view but are prone to interference in complex environments [5,6]. Thermal imaging techniques, which identify fire sources by detecting thermal radiation, are adaptable but their high cost and sensitivity to changes in ambient temperature limit their widespread use [7,8]. Despite their capacity to discern smoke and flames, conventional techniques are hindered by limitations such as inadequate sensitivity, a high rate of false alarms, and poor environmental adaptability.

In recent years, the advent of deep learning techniques has yielded novel solutions for smoke and flame detection. Deep learning frameworks, including Faster R-CNN [9] and SSD [10] (Single Shot MultiBox Detector), demonstrate efficacy in target detection. However, they exhibit deficiencies in real-time and environment adaptation, rendering them ill suited for fire monitoring scenarios necessitating a rapid response. To address these limitations, the YOLO [11,12] (You Only Look Once) family of algorithms was developed, garnering significant attention in real-world applications due to its superior detection speed and high accuracy rate. With each iteration from YOLOv5 to YOLOv11, the algorithm has been refined to enhance the practicality of flame and smoke detection in natural scenes, Specifically, shortcomings of previous versions—such as limited efficacy in low-contrast conditions, challenges in detecting small-scale targets, and a relatively high false-positive rate in complex backgrounds—have been addressed, establishing YOLOv11 as the preferred solution for fire monitoring. Nevertheless, substantial challenges persist in complex natural environments, particularly in scenarios characterized by intricate backgrounds and intertwined smoke and flames, which frequently exhibit suboptimal detection accuracy and elevated false alarm rates [13,14]. Researchers have gradually introduced constrained strategies and multiscale feature fusion strategies to improve detection accuracy. The optimization of the feature extraction layer, along with the incorporation of an attention mechanism, has demonstrated enhanced focalization on critical regions, leading to substantial improvements in performance in complex scenes [15–18]. Furthermore, techniques such as data enhancement and migration learning have been shown to enhance the model's adaptability and robustness in different environments [19].

Despite progress in smoke and flame detection, numerous technical challenges remain, including small object detection, occlusion issues in dynamic backgrounds, and

the recognition of targets with ambiguous boundaries. Addressing these problems, this paper proposes a lightweight deep learning model, SF-YOLO (Smoke and Fire-YOLO), suitable for smoke and flame detection in natural scenes. The SF-YOLO model is based on YOLOv11′s C3k2 module [20] and incorporates a dual-path residual attention mechanism to enhance the accuracy of small object detection. As explained in the next section, by employing a detection head with an embedded attention mechanism, the response of unobscured regions is enhanced, compensating for the feature loss in occluded regions, thus improving the detection accuracy of occluded targets. Finally, a loss function based on dynamic adjustment and intelligent focusing is utilized to enhance loss computation in boundary regions, enabling the model to more accurately recognize targets with ambiguous boundaries. Based on these improvements, the proposed algorithms are experimentally validated using the self-generated dataset S-Firedata and the publicly available dataset M^4SFWD [21]. The results demonstrate that the algorithm effectively detects and locates smoke and flame targets in natural scenes.

2. Materials and Methods

To address the demand for smoke and flame detection in natural scenes, this paper proposes a lightweight deep learning algorithm for smoke and flame detection. The overall technical process of the algorithm is illustrated in Figure 1. First, the input data are frame images extracted from real-time video captured by cameras, preprocessed to standardize image dimensions and enhance contrast. The preprocessed images are then subjected to feature extraction using the improved YOLOv11 network model, SF-YOLO. At this stage, the C3k2 module, which integrates a dual-path residual attention mechanism, effectively enhances the recognition capability for blurred boundaries and occluded targets. Following a series of multi-scale concatenation and upsampling operations, the extracted features are transmitted to a detection head with an embedded attention mechanism. This module dynamically adjusts the focus areas and employs an intelligent focusing loss function to optimize the training process of the model, thereby addressing the target detection issues in complex backgrounds. After a series of processing steps, the model outputs the final smoke and flame detection results. These results include the position, category, and confidence level of the targets, providing reliable data support for real-time fire monitoring and emergency response.

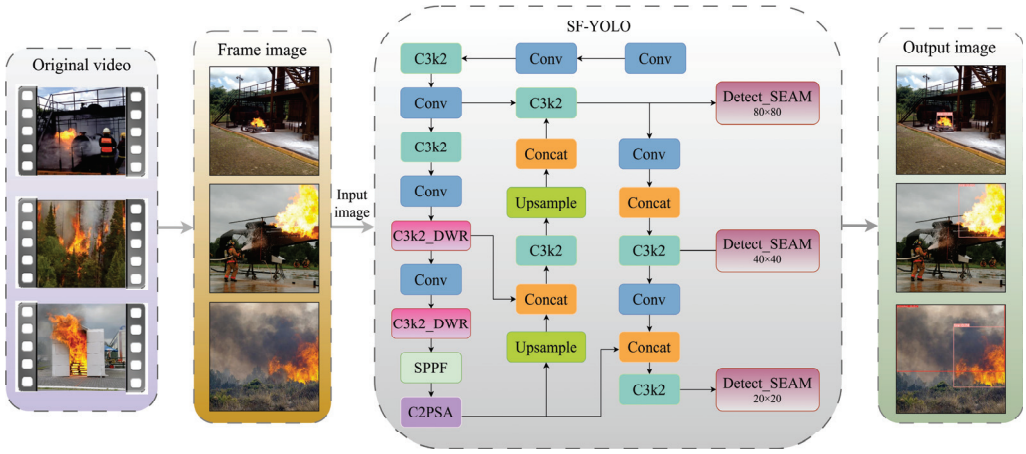


Figure 1. Overall technical flowchart of the algorithm in this paper.

2.1. C3k2 Module Based on Dual-Path Residual Attention Mechanism

In this paper, we introduce the C3k2 module based on DWR [22] (Dilation-Wise Residual) with the aim of enhancing the efficiency of real-time feature extraction using a two-step feature extraction approach. The DWR mechanism, respectively, utilizes regional

residualization and semantic residualization strategies in high-level and low-level networks to adapt to the requirements of different receptive fields. The term "strong semantic information" refers to the incorporation of high-quality labeled data to derive meaningful features that enhance the accuracy and robustness of detection. These semantics are derived from well-labeled datasets that provide a rich source of context and enable more precise pixel classification, thereby improving the overall effectiveness of the model.

The C3k2-DWR module comprises two submodules: DWR and SIR (Simple Inverted Residual). This architecture is designed to optimize feature extraction and improve the overall performance of the model. The structure of the C3k2-DWR module is illustrated in Figure 2, showcasing its innovative design and the integration of the DWR mechanism to achieve superior feature extraction capabilities. By leveraging this dual approach, the C3k2-DWR module aims to deliver enhanced accuracy and efficiency in real-time applications.

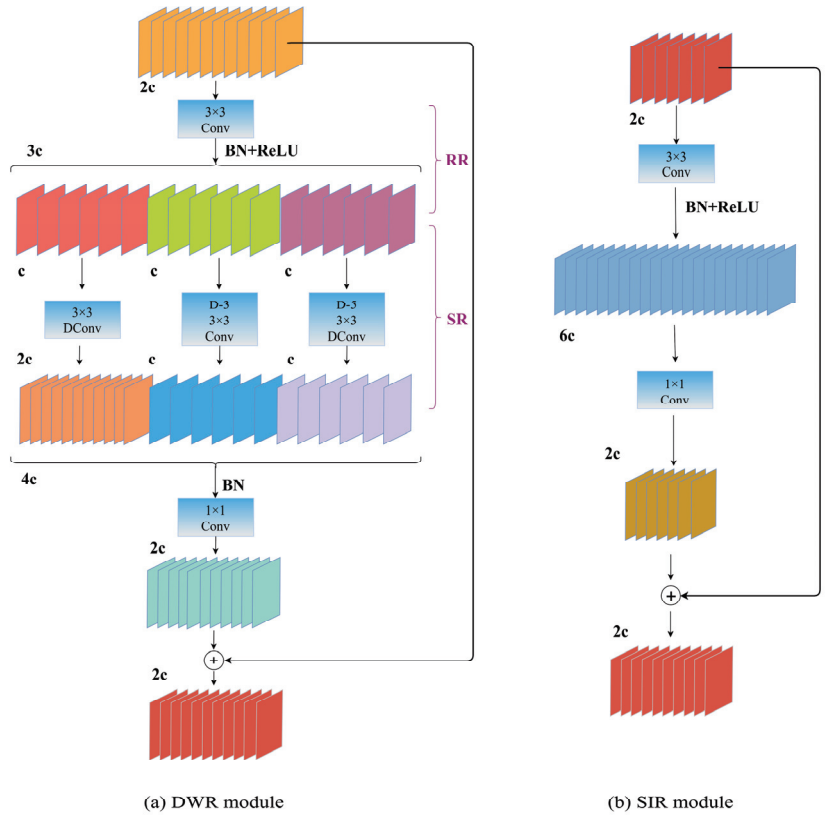


Figure 2. The two modules of C3k2_DWR.

The operation flow of the DWR module depicted in Figure 2a is as follows: The initial features are first extracted by a 3 × 3 convolutional layer, followed by batch normalization (BN) and ReLU activation to enhance feature representation. Subsequently, the data are directed into multiple dilated depth convolutional paths with different dilation rates (e.g., 3 × 3 dilated convolutional layers of D-3 and D-5) to extend the receptive field and capture a broader range of contextual information [23]. The output feature maps from these paths are then combined through a concatenation (Concat) operation. These combined feature maps are further processed by batch normalization and 1 × 1 convolutional layers. Finally, the module's output is obtained by summing the initial input with the processed results.

As illustrated in Figure 2b, the data undergoes initial processing through a 3 × 3 convolutional layer in the SIR module, followed by batch normalization and a rectified linear unit (ReLU) activation function. The feature map is then further processed through a 1 × 1 convolutional layer, culminating in the summation of this layer with the initial input to yield the module's final output.

Finally, the feature maps of the DWR and SIR modules are upsampled using strong semantic information and spliced with the lower-stage feature maps of flow. The final output predicted feature maps, C3k2-DWR, are shown in Equation (1) to achieve the final smoke and flame prediction.

$$\begin{cases} F_{output} = Upsample(Concat(F_{DWR}, F_{SIR})) \\ F_{C2f-DWR} = Concat(F_{output}, F_{low}) \end{cases} \quad (1)$$

2.2. Target Detection Header Framework with Embedded Attention Mechanism

SEAMHead [24] (Separated and Enhancement Attention Module Head) is a sophisticated deep learning model designed to enhance the accuracy and robustness of smoke flame detection. It utilizes a multi-head attention mechanism and a deeply separable convolution strategy to emphasize important flame regions, reduce background interference, and enhance feature representation through residual linking and information fusion. This approach significantly improves the model's performance under occlusion and in complex backgrounds.

The SEAM module is composed of three CSMM (Channel and Spatial Mixing Module) blocks, with block sizes of 6, 7, and 8, as illustrated in Figure 3. Each CSMM block processes multi-scale features through patch embedding to effectively learn the correlation between spatial dimensions and channels. The following details the specific process.

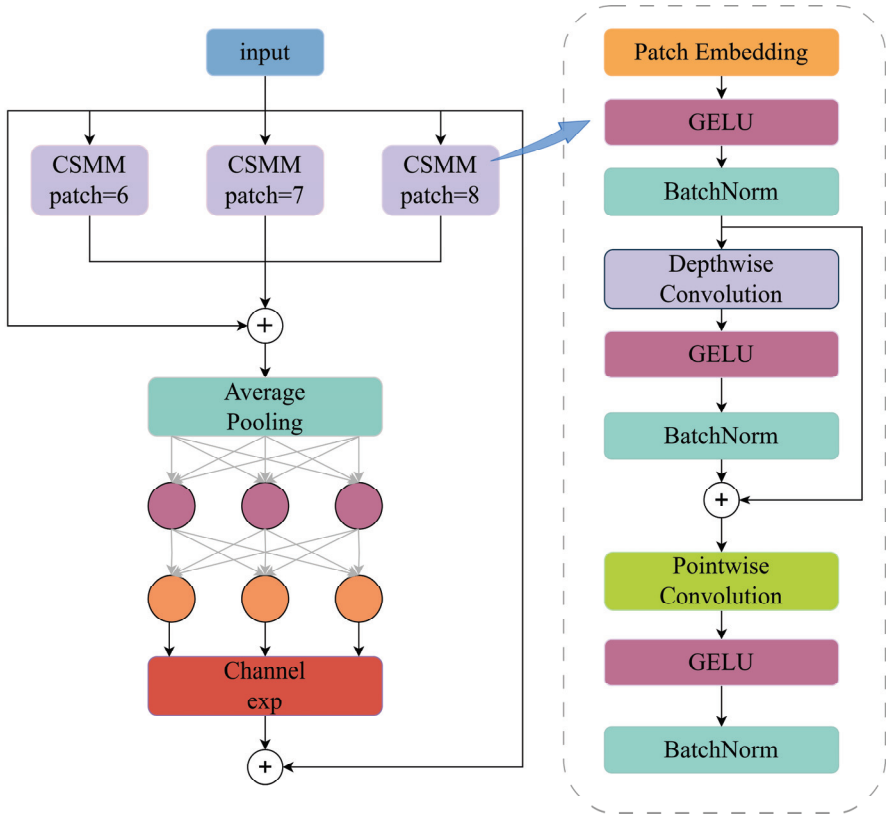


Figure 3. SEAMHead's Fully Connected Network Architecture.

First, the CSMM module utilizes the patch embedding technique to process multiscale features, effectively capturing the correlation between spatial dimensions and channels. This process consists of GELU (Gaussian Error Linear Unit) activation function and batch normalization (BatchNorm), followed by feature extraction through dilated depth convolution, and then further GELU and BatchNorm processing. Finally, feature fusion is achieved through Residual Connection and Pointwise Convolution. The aforementioned sequence

of operations is illustrated in Equation (2), where Patchi denotes the patch extracted from the input feature map.

$$Z_i = GELU(BatchNorm(DilationConv(Patch_i))) + Patch_i \tag{2}$$

Subsequently, the CSMM-processed feature maps undergo average pooling, which captures features at varying scales. These features are then integrated through pointwise convolution with the outputs of various depth convolutions (e.g., 3 × 3 and 5 × 5) to enhance their inter-channel connectivity. The output of the Fully Connected Layer (FCL) is then processed by an exponential function to extend the range of values from [0, 1] to [1, e]. As demonstrated in Equation (3), this exponential normalization method establishes a monotonic mapping relation, enhancing the tolerance of the results to positional errors [25].

$$\begin{cases} F = Conv3x3(AvgPool(Z)) + Conv5x5(AvgPool(Z)) \\ O = e^F \end{cases} \tag{3}$$

In conclusion, the output of the SEAM module is multiplied by the original features as attentional weights. This process enables the more efficient management of occluded smoke flame targets.

2.3. A Dual-Channel Loss Function Based on Dynamic Adjustment and Intelligent Focusing

During the target detection process, the coverage of the anchor frame and the actual frame is determined by calculating the intersection over union (IoU). This method effectively eliminates the interference caused by differences in the size of bounding boxes, enabling the model to balance the learning of large and small objects when using 1-IoU as the binary bounding region (BBR) loss. The functional definition of the IoU loss [26] is shown in Equation (4).

$$\mathcal{L}_{IoU} = 1 - \text{IoU} = 1 - \frac{W_i H_i}{S} \tag{4}$$

The SIoU (SCYLLA-IoU) loss function provides a comprehensive approach to loss calculation by incorporating angular differences in anchor frames, comparing the widths and heights of predictor and target frames, and calculating the distance between centroids [27]. WIoU (Wise-IoU) [28] introduces a dynamic non-monotonic focusing mechanism that dynamically adjusts the gradient gain by evaluating the anomaly degree (outlier degree) of anchor frames. This mechanism enables the model to prioritize ordinary quality anchor frames, thereby reducing the competitiveness of high-quality anchor frames and mitigating the negative impact of low-quality samples. This paper employs the SIoU loss based on the dynamic non-monotonic focusing mechanism, denoted as W-SIoU, as depicted in Figure 4.

The structural flow of W-SIoU begins with the acquisition of the prediction frame and the real frame of the model. Subsequently, the WIoU loss and SIoU loss are calculated. WIoU is based on the distance between the prediction frame and the real frame. Next, distance attention is calculated to enhance attention to the ordinary quality samples. This process reduces large or harmful gradients in extreme samples through a reasonable gradient gain allocation strategy. The formula is shown in Equation (5).

$$\mathcal{L}_{WIoU} = \frac{\sum_{i=1}^{N} w_1 \cdot IoU_1}{\sum_{i=1}^{N} w_1} \tag{5}$$

where w_i is the weight of the i-th sample, IoU_i is the IoU value of the i-th sample, and N is the total number of samples. The process structure of the SIoU function is as follows: First, the vector angle between the predicted box and the real box is calculated to reduce the phenomenon of the predicted box's "wandering". Then, the distance loss is redefined so

that the contribution of the distance loss is reduced at small angles and the shape alignment of the bounding box is optimized. Subsequently, the distance loss is redefined such that its contribution is reduced when the angle is small and the shape alignment of the bounding box is optimized. Finally, the intersection over union (IoU) loss between the prediction frame and the real frame is calculated by combining the traditional IoU loss. The calculation formula is shown in Equation (6).

$$\begin{cases} x = \frac{c_h}{\sigma} = \sin(\alpha) \\ \sigma = \sqrt{\left(b_{c_x}^{gt} - b_{c_x}\right)^2 + \left(b_{c_y}^{gt} - b_{c_y}\right)^2} \\ c_h = \max\left(b_y^{gt}, b_y\right) - \min\left(b_y^{gt}, b_y\right) \end{cases} \tag{6}$$

where ch denotes the length of the vertical edge, σ denotes the distance between the two frame centroids (b_{cx}, b_{cy}), the coordinates of the predicted frame centroids are denoted by ($b^{gt}{}_{cx}$, $b^{gt}{}_{cy}$), and the coordinates of the real frame centroids are denoted by ($b^{gt}{}_{cx}$, $b^{gt}{}_{cy}$).

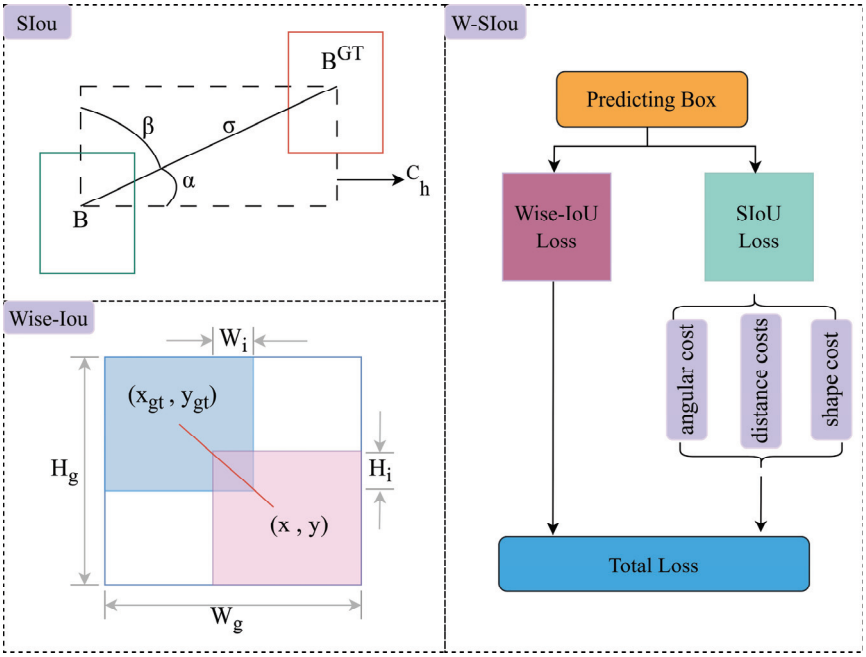


Figure 4. W-SIoU Schematic.

Subsequently, WIoU and SIoU losses are weighted and combined to obtain the final integrated loss function, W-SIoU. This function optimizes the positioning accuracy and shape alignment of the bounding box, thereby improving target detection accuracy. Concurrently, a multifaceted evaluation process mitigates the influence of outliers, enhancing the model's resilience in complex scenarios. This approach enables more precise detection of smoke and flame targets with indistinct boundaries in natural environments.

3. Results

3.1. Experimental Data

To carry out the experimental validation of the algorithms in this paper, we produced the smoke and fire dataset S-Firedata (Smoke and Fire data), which contains 240 images from the publicly available BoWFireDataset [29], the D-Fire dataset [30], which contains 6860 images; the Flame fire dataset [31], which contains 802 images; the fire-8 flame dataset, which contains 919 images; and the video framing homemade data, which contains 1820 images, for a total of 10,641 images. It contains two classes: smoke and fire. Of these,

2855 images contain exclusively smoke targets, 2447 exclusively flame targets, 2672 contain both smoke and flame targets, and 2667 do not contain smoke or flame targets. For the experimental testing phase, 7447 images were designated for the training set (train), 1597 images for the validation set (val), and 1597 images for the test set (test). Additionally, the open-source dataset M^{4}SFWD was utilized as a control experiment. It was constructed with Unreal Engine 5 to build eight virtual fire scenes with different terrains, containing a total of 3974 images. The partitioning of the data are shown in Table 1 below.

Table 1. Composition of the smoke and fire data dataset.

Dataset			
Parameters	No. of images	Parameters	No. of images
BoWFireDataset	240	Fire	2447
D-Fire dataset	6860	Smoke	2855
the Flame fire	802	Fire and Smoke	2672
fire-8	919	No smoke flames	2667
homemade data	1820	Train	7447
S-Firedata(all)	10,641	Val	1597
M^{4}SFWD	3947	Test	1597

The experimental hardware environment includes Windows 11 as the operating system, an Intel Core i7 13700F@5.40GHz CPU, 16 GB of memory, and an NVIDIA GeForce RTX 3060 graphics card with 20 GB of video memory. The software environment consists of Python 3.8.18 and the Pytorch-GPU 2.1.0 framework, paired with CUDA version 11.8. The model pretraining weights are YOLOv11m.pt, with the training period set to 100 epochs and batch processing configured with a batch size of four.

In the YOLO series of models, the evaluation indexes for model precision include Precision, Recall, F_1 Score, and Mean Average Precision (mAP) [32]. The primary performance reference index in this paper is the F_1 Score, with mAP50 serving as the secondary reference index. The F_1 Score, which calculates the harmonic mean of precision and recall, offers a comprehensive evaluation of a model's performance across these two dimensions. The mAP50 metric signifies the value of mAP when the intersection over union (IoU) ratio attains 0.5, while mAP50-95 indicates the value of IoU from 50% to 95% in increments of 5%. mAP50-95 signifies the average mAP value when IoU increases from 50% to 95% in steps of 5%. An increase in the mAP value corresponds to an enhancement in the model's overall accuracy [33]. Evaluation indexes are computed by Formulas (7) and (8).

$$\begin{cases} P = \frac{TP}{TP+FP} \\ R = \frac{TP}{TP+FN} \\ F_1 = \frac{2 \cdot P \cdot R}{P+R} \end{cases} \tag{7}$$

$$\begin{cases} AP = \int_0^1 PRdr \\ mAP = \frac{1}{n}\sum_{i=1}^{n} APi \end{cases} \tag{8}$$

In machine learning models, TP (true positives) denotes the number of samples correctly predicted as positive categories by the model. FP (false positives) represents the number of samples incorrectly predicted as positive categories. FN (false negatives) indicates the number of samples incorrectly predicted as negative categories. AP (average precision) refers to the average of the area under the precision–recall curve for each category. The mean average precision (mAP) is obtained by averaging the AP values across all categories.

3.2. Comparison of Detection Performance of Different Models

To evaluate the performance of different object detection algorithms, this paper selects several popular object detection models for comparative experiments. These models include Faster-RCNN, which addresses high computational and structural complexity issues of R-CNN, Centernet with anchor-free detection [34], Huawei's efficient real-time object detector Gold-Yolo [35], and the YOLO series models YOLOv7 [36], YOLOv8, and YOLOv11. All experiments were conducted using the self-generated dataset S-Firedata, with identical numbers of images in the test set, validation set, and training set. Each model, including the improved model proposed in this paper, was trained 100 times using a fixed batch size of four, and all other hyperparameter settings were maintained identically. The precision (P), recall (R), F_1 Score, mean average precision (mAP), and other quantitative evaluation metrics of these models were compared, as shown in Table 2.

Table 2. Comparison of the accuracy of different target detection models for smoke flame detection against the S-Firedata dataset.

Model Structure	P (%)	R (%)	F_1 Score	mAP50 (%)	mAP50-95 (%)	Params (M)
Centernet	66.5	59.2	62.64	58.6	32.8	40
Faster-RCNN	53.4	58.2	55.68	52.5	28.5	40
Gold-Yolo	69.3	59.6	64.09	64.0	36.8	59
YOLOv7	67.4	59.0	62.92	56.5	27.3	36
YOLOv8	72.3	59.3	65.16	65.1	37.8	26
YOLOv11	72.5	61.8	66.72	66.7	38.4	20
SF-YOLO (Ours)	**77.9**	**64.8**	**70.75**	**69.2**	**40.8**	**14**

As illustrated in Table 2, the accuracy metric of YOLOv11 is comparable to that of Gold-Yolo, with an mAP50 reaching 66.7%. This mAP50 score is significantly superior to other popular object detection models, such as Faster-RCNN and Centernet. Therefore, we chose YOLOv11 as the baseline network for algorithm enhancement, which not only improves performance but also significantly reduces the number of parameters. Additionally, in terms of model parameter storage capacity, our SF-YOLO model has the smallest footprint among all the compared models.

3.3. Ablation Experiments on Different Modules

To verify the impact of different modules on the performance of deep learning models, this paper conducts a series of ablation experiments on the self-made dataset S-Firedata, and the comparisons of accuracy metrics are shown in Table 3.

Table 3. Results of ablation experiments of deep neural network models on the S-Firedata dataset.

Baseline	IoU	C3k2-DWR	SEAMHead	Precision (%)	F_1 Score	mAP50 (%)	mAP50-95 (%)
YOLOv11	CIoU	×	×	72.5	66.72	66.7	38.4
YOLOv11	W-SIoU	×	×	74.3	69.05	67.6	39.2
YOLOv11	CIoU	✔	×	75.4	68.82	67.2	39.1
YOLOv11	CIoU	×	✔	75.9	68.55	67.6	39.3
YOLOv11	CIoU	✔	✔	76.5	69.04	67.9	39.5
YOLOv11	W-SIoU	✔	✔	76.9	69.14	68.6	39.8
YOLOv11 *	W-SIoU	✔	✔	**77.9**	**70.75**	**69.2**	**40.8**

×: This module is not used in the algorithm, ✔: Use this module in algorithms, *: The algorithm performed transfer learning.

As illustrated in Table 3, the incorporation of a single enhanced structure in various datasets does not result in a substantial enhancement of model accuracy. When the advanced C3k2 module and the detection head are integrated concurrently, the model

demonstrates a modest improvement in accuracy compared to the baseline. However, the efficacy of target detection does not improve largely. After substituting the original loss function with the self-developed hybrid loss function W-SIoU, the model's accuracy is further enhanced. Additionally, the self-developed W-SIoU loss function exhibits a substantial reduction in parameters and computational demands compared to the CIoU parameter number. The enhancement of the self-developed W-SIoU loss function, in comparison to the CIoU parameter number and computational volume, has been significantly reduced. The trained model is more lightweight, and the lightweight model in this paper is based on YOLOv11 by adding C3k2-DWR and replacing the detector head SEAMHead. The W-SIoU loss function is replaced after migration learning, and the model is finally trained, which further improves the F_1 Score. The accuracy rate exceeds the benchmark, and all the metrics are significantly improved compared to the benchmark.

SF-YOLO outperforms other models in mAP50 by reliably identifying and classifying objects even with moderate localization deviations. Its advanced architectural design further elevates performance in the more stringent mAP50-95 metric by ensuring precise object localization and robust feature extraction. This dual strength underscores the model's comprehensive and balanced approach to object detection.

In order to verify the effectiveness of the paper's algorithm, the SF-YOLO model was implemented in the open-source dataset M^4SFWD, and the ablation experiments were conducted to assess the accuracy of various comparisons. As shown in Table 4, the SF-YOLO model exhibited a substantial enhancement in the accuracy of various precision indexes when compared to the benchmark accuracy of the open dataset. This finding further substantiates the efficacy of the SF-YOLO algorithm in automatic smoke and flame detection.

Table 4. Experimental results of deep neural network model ablation on M^4SFWD dataset.

	IoU	C3k2-DWR	SEAMHead	P (%)	F_1 Score	mAP50 (%)	mAP50-95 (%)
1	CIoU	×	×	83.2	82.19	86.4	49.5
2	W-SIoU	×	×	83.3	82.24	86.5	50.1
3	CIoU	✔	×	83.6	82.23	86.5	50.1
4	CIoU	×	✔	83.6	82.33	86.6	50.1
5	CIoU	✔	✔	83.9	82.32	86.5	50.3
SF-YOLO	W-SIoU	✔	✔	**84.9**	**83.73**	**87.1**	**50.9**

4. Discussion

To more intuitively demonstrate the effectiveness of the proposed method, Figures 5–9 illustrate the target detection performance of various deep learning models on remote sensing fire targets, multiple targets, small targets, targets in dark environments, occluded targets, and mixed targets with similar objects. Additionally, for sunset targets that have a similar color to flames, the proposed method can effectively distinguish them, as shown in Figure 10.

Figures 5 and 6 illustrate that the Centernet, YOLOv7, and Faster-RCNN models exhibit higher leakage and false detection rates. In contrast, the Gold-YOLO, YOLOv8, and YOLOv11 models display a lower incidence of these issues. However, these models tend to have a higher false detection rate under occlusion conditions, as shown in Figure 9. In comparison, the proposed SF-YOLO model demonstrates superior robustness and accuracy in these scenarios, particularly when dealing with dynamic backgrounds and complex occlusions. This highlights the effectiveness of the SF-YOLO model over other popular models.

A recent study documented a significant forest fire in California [37,38]. The SF-YOLO model was employed to analyze the satellite images released by Maxar, as depicted in

Figure 11. This analysis demonstrated the model's efficacy in detecting both small flame targets and large flame patches in satellite imagery. This finding further substantiates the model's effectiveness and provides a quantitative basis for the conclusions drawn in this paper.

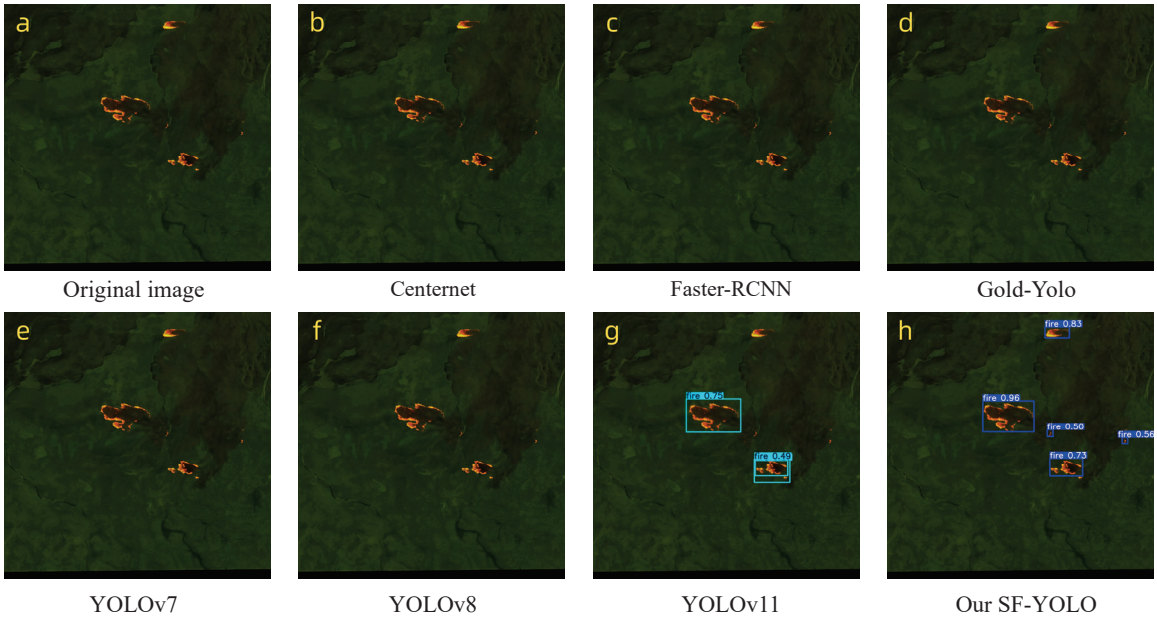


Figure 5. Smoke flame detection effectiveness of different deep neural network models in the case of remote sensing fire targets. With the exception of YOLOv11 and SF-YOLO, none of the models detected the target. (**a**) is the original image; (**b**) is the Centernet detection effect image; (**c**) is the Faster-RCNN detection effect; (**d**) is the Gold-Yolo detection effect; (**e**) is the YOLOv7 detection effect; (**f**) is the YOLOv8 detection effect; (**g**) is the YOLOv11 detection effect; (**h**) is the detection effect of the proposed algorithm.

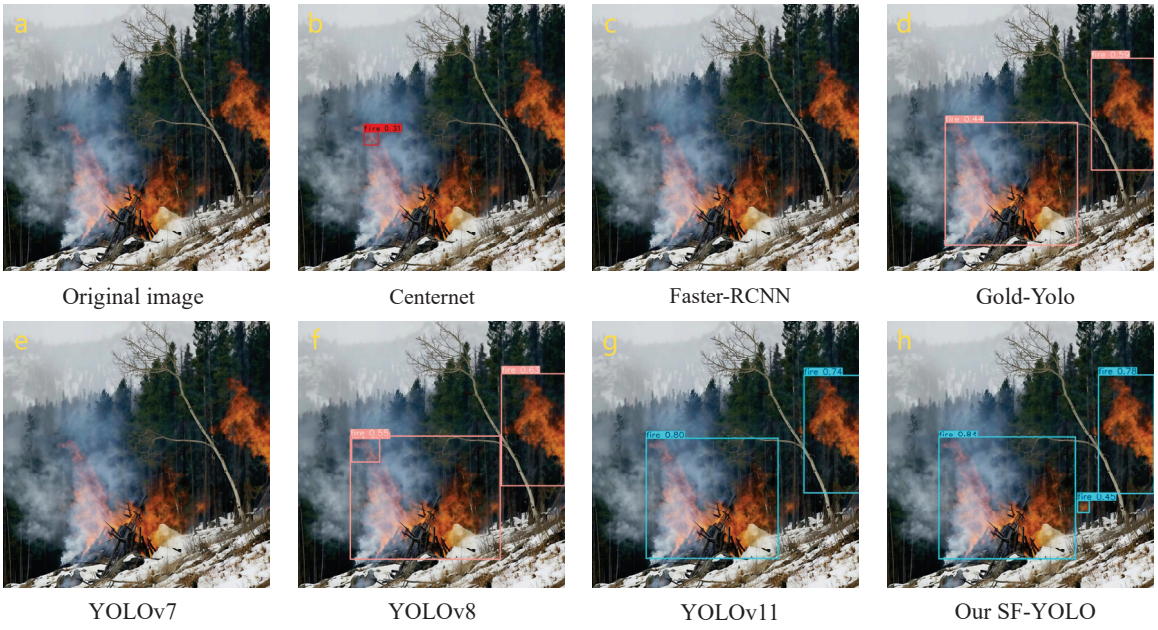


Figure 6. Effectiveness of different deep neural network models for smoke flame detection in multi-target situations. Centernet can only detect a portion of the targets and none of the models except SF-YOLO detect small targets in the image. (**a**) is the original image; (**b**) is the Centernet detection effect image; (**c**) is the Faster-RCNN detection effect; (**d**) is the Gold-Yolo detection effect; (**e**) is the YOLOv7 detection effect; (**f**) is the YOLOv8 detection effect; (**g**) is the YOLOv11 detection effect; (**h**) is the detection effect of the proposed algorithm.

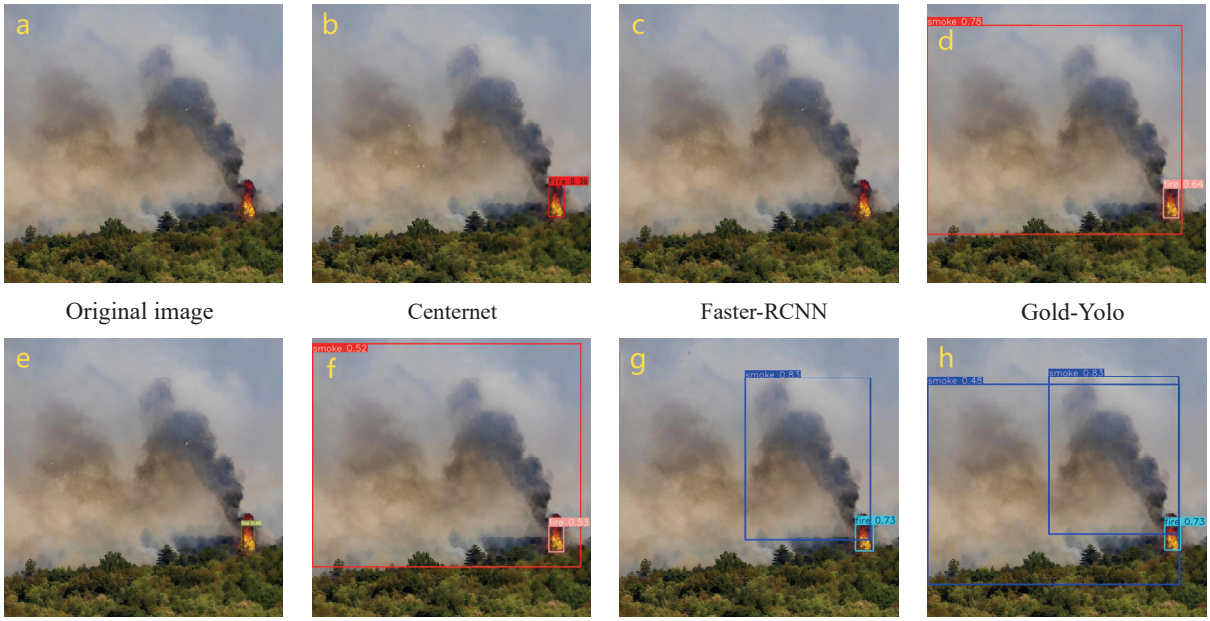


Figure 7. Effectiveness of different deep neural network models for smoke flame detection in the case of including small targets. Only SF-YOLO successfully recognizes all targets on the image. (**a**) is the original image; (**b**) is the Centernet detection effect image; (**c**) is the Faster-RCNN detection effect; (**d**) is the Gold-Yolo detection effect; (**e**) is the YOLOv7 detection effect; (**f**) is the YOLOv8 detection effect; (**g**) is the YOLOv11 detection effect; (**h**) is the detection effect of the proposed algorithm.

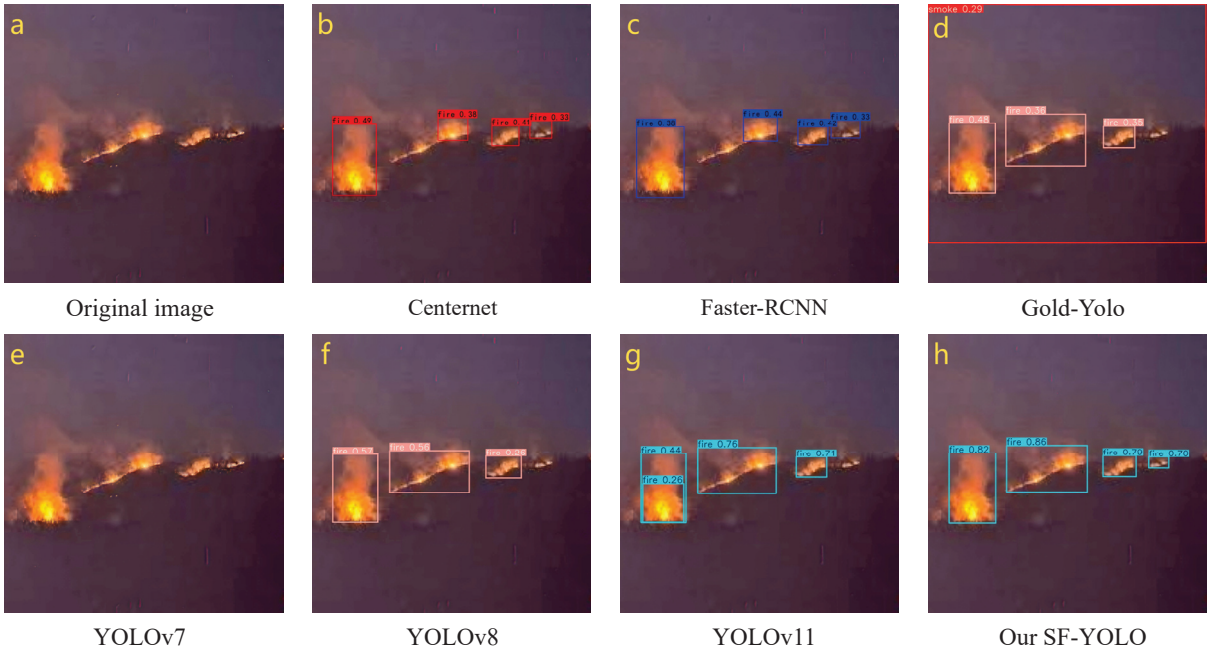


Figure 8. Effectiveness of different deep neural network models for smoke flame detection in situations containing targets in dark environments. Centernet and Faster-RCNN have leakage detection, and all other models have lower detection accuracy than SF-YOLO. (**a**) is the original image; (**b**) is the Centernet detection effect image; (**c**) is the Faster-RCNN detection effect; (**d**) is the Gold-Yolo detection effect; (**e**) is the YOLOv7 detection effect; (**f**) is the YOLOv8 detection effect; (**g**) is the YOLOv11 detection effect; (**h**) is the detection effect of the proposed algorithm.

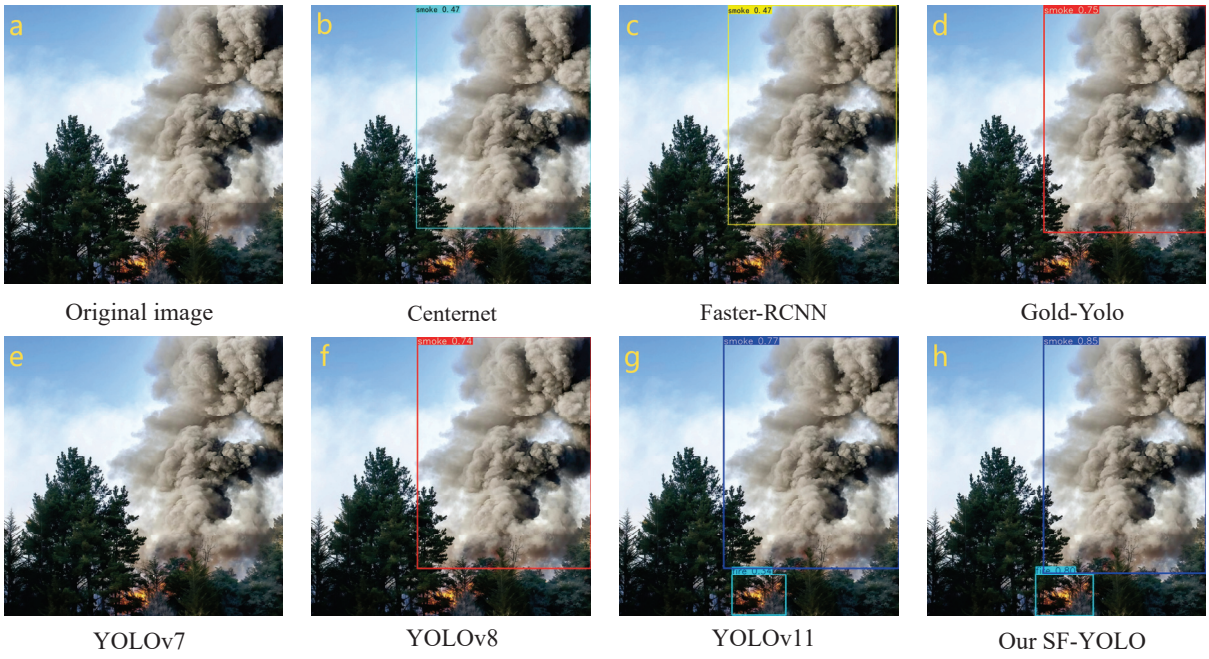


Figure 9. Smoke flame detection effectiveness of different deep neural network models in the case of including occluded targets. With the exception of YOLOv11 and SF-YOLO, all models were designed to detect flame targets other than those obscured by foliage. (**a**) is the original image; (**b**) is the Centernet detection effect image; (**c**) is the Faster-RCNN detection effect; (**d**) is the Gold-Yolo detection effect; (**e**) is the YOLOv7 detection effect; (**f**) is the YOLOv8 detection effect; (**g**) is the YOLOv11 detection effect; (**h**) is the detection effect of the proposed algorithm.

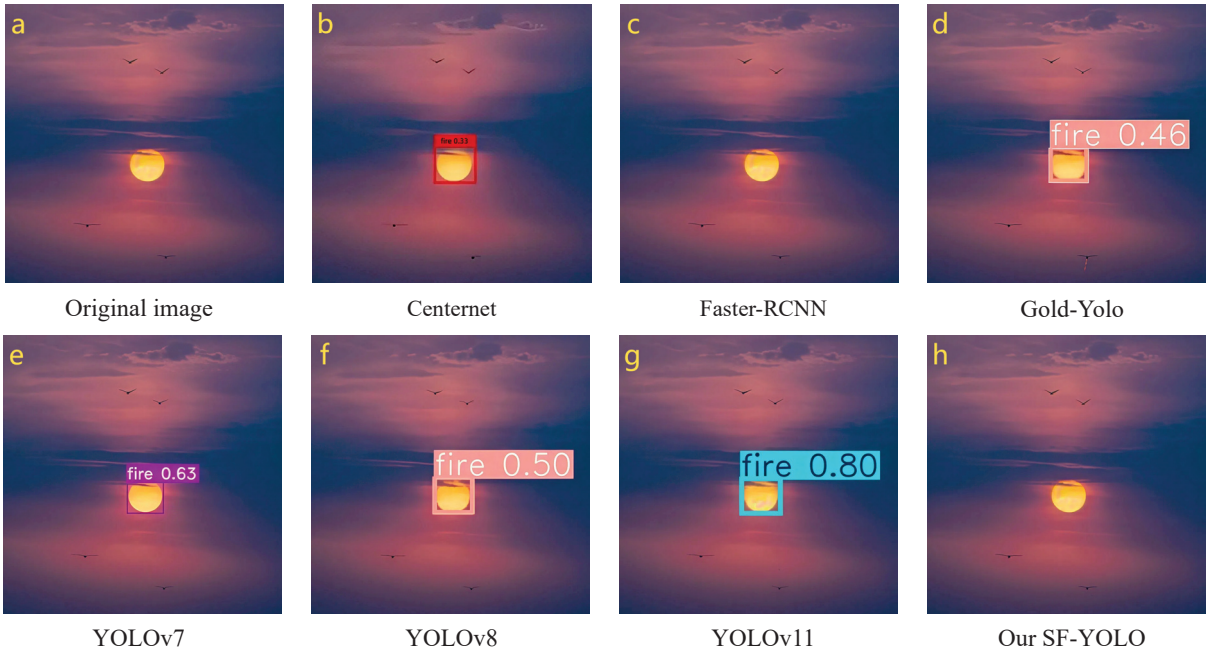


Figure 10. Effectiveness of different deep neural network models for smoke flame detection in the case of including fire-like targets. Only SF-YOLO and Faster-RCNN succeeded in identifying the obfuscated target. (**a**) is the original image; (**b**) is the Centernet detection effect image; (**c**) is the Faster-RCNN detection effect; (**d**) is the Gold-Yolo detection effect; (**e**) is the YOLOv7 detection effect; (**f**) is the YOLOv8 detection effect; (**g**) is the YOLOv11 detection effect; (**h**) is the detection effect of the proposed algorithm.



Figure 11. Detection effectiveness of the SF-YOLO algorithm in the Los Angeles Hill Fire. The algorithm in this paper accurately detects and identifies the scattered small fires in the graph with a confidence level of about 40% for the tiny targets.

5. Conclusions

To address the limitations of traditional methods in detecting targets with significant morphological alterations, ambiguous targets with unclear boundaries, and occluded targets in dynamic backgrounds, this paper proposes a lightweight deep learning model, SF-YOLO, for detecting smoke and flame in natural scenes. Our SF-YOLO model incorporates the C3k2 module based on a dual-path residual attention mechanism, a target detection head framework with an embedded attention mechanism, and a loss function based on dynamic adjustment and intelligent focusing. This combination enhances the response of unobscured regions, compensates for feature loss in occluded regions, and addresses occlusion issues in dynamic backgrounds. A series of comparative experiments were conducted on the public dataset M^4SFWD and the self-made dataset S-Firedata to evaluate the performance of the algorithms. The experimental findings demonstrate that the proposed algorithm enhances the detection of non-rigid moving targets under conditions of target occlusion and boundary feature blurring, among other scenarios, in comparison with the original YOLOv11. The SF-YOLO model exhibits higher accuracy and robustness in detecting smoke and flame in natural scenes. The F_1 Score metrics for the two distinct datasets, M^4SFWD and S-Firedata, show significant improvements of 1.54% and 4.03%, respectively, and the SF-YOLO algorithm also has good detection results in images from the Los Angeles Hill Fire. Additionally, the proposed algorithm is more compact and lightweight compared to the original YOLOv11 model. Consequently, the research presented in this paper provides a practical solution for the automatic detection of smoke and flame in natural scenes, thereby enhancing the real-time accuracy of fire warning systems.

Building on the current work, future research should address the critical link between wildfire occurrence distributions and image characteristics by integrating environmental covariates (e.g., landscape topology, forest compositions, and fuel moisture content) into model training. This could involve synthesizing multi-spectral datasets that align fire-prone geographic profiles with their spectral–spatial signatures, mitigating biases arising from underrepresented terrains or fuel types. Additionally, developing domain adaptation frameworks to account for seasonal variations in vegetation and aspect-driven illumination changes would enhance generalizability.

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

Funding: This research was funded by the National Natural Science Foundation of China, grant numbers 42271342 and 42471463 and the Priority Academic Program Development of Jiangsu Higher Education Institutions, grant number 164320H116.

Institutional Review Board Statement: Not applicable for studies not involving humans or animals.

Informed Consent Statement: Not applicable for studies not involving humans.

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 authors sincerely thank the School of Geography Science at Nanjing Normal University for their support throughout the entire project. We also wish to express our gratitude to the Key Laboratory of Virtual Geographic Environment, Ministry of Education (Nanjing Normal University), for providing the necessary resources, hardware, and technical assistance, which greatly contributed to our research efforts.

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

References

1. Suhardono, S.; Fitria, L.; Suryawan, I.W.K.; Septiariva, I.Y.; Mulyana, R.; Sari, M.M.; Ulhasanah, N.; Prayogo, W. Human activities and forest fires in Indonesia: An analysis of the Bromo incident and implications for conservation tourism. *Trees For. People* **2024**, *15*, 100509. [CrossRef]
2. Hu, Q.; Zhang, R.; Qian, X.; Yuan, M.; Li, P. Explosion Shock Dynamics and Hazards in Complex Civil Buildings: A Case Study of a Severe Fuel Explosion Accident in Yinchuan, China. *Fire* **2024**, *7*, 310. [CrossRef]
3. Hsu, T.W.; Pare, S.; Meena, M.S.; Jain, D.K.; Li, D.L.; Saxena, A.; Prasad, M.; Lin, C.T. An early flame detection system based on image block threshold selection using knowledge of local and global feature analysis. *Sustainability* **2020**, *12*, 8899. [CrossRef]
4. Ricci, F.; Mariani, F. Advanced Flame front Detection in Combustion Processes Using Autoencoder Approach. *Energies* **2024**, *17*, 1759. [CrossRef]
5. Barmpoutis, P.; Papaioannou, P.; Dimitropoulos, K.; Grammalidis, N. A review on early forest fire detection systems using optical remote sensing. *Sensors* **2020**, *20*, 6442. [CrossRef] [PubMed]
6. Si, R. Recent Advancements in Flame Sensors. In *IET Conference Proceedings CP901*; The Institution of Engineering and Technology: Stevenage, UK, 2024; Volume 2024, pp. 43–47.
7. Sousa, M.J.; Moutinho, A.; Almeida, M. Thermal infrared sensing for near real-time data-driven fire detection and monitoring systems. *Sensors* **2020**, *20*, 6803. [CrossRef]
8. Zheng, F.; Sun, G.; Suo, Y.; Ma, H.; Feng, T. Research on Through-Flame Imaging Using Mid-Wave Infrared Camera Based on Flame Filter. *Sensors* **2024**, *24*, 6696. [CrossRef]
9. Cheng, B.; Wei, Y.; Shi, H.; Feris, R.; Xiong, J.; Huang, T. Revisiting Rcnn: On Awakening the Classification Power of Faster Rcnn. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 453–468.
10. Liu, S.; Huang, L.; Zhao, Y.; Wu, X. Lightweight Single Shot Multi-Box Detector: A fabric defect detection algorithm incorporating parallel dilated convolution and dual channel attention. *Text. Res. J.* **2024**, *94*, 209–224. [CrossRef]
11. Jiang, P.; Ergu, D.; Liu, F.; Cai, Y.; Ma, B. A Review of Yolo algorithm developments. *Procedia Comput. Sci.* **2022**, *199*, 1066–1073. [CrossRef]
12. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
13. Cao, X.; Su, Y.; Geng, X.; Wang, Y. YOLO-SF: YOLO for fire segmentation detection. *IEEE Access* **2023**, *11*, 111079–111092. [CrossRef]
14. Wang, D.; Qian, Y.; Lu, J.; Wang, P.; Hu, Z.; Chai, Y. Fs-yolo: Fire-smoke detection based on improved YOLOv7. *Multimed. Syst.* **2024**, *30*, 215. [CrossRef]
15. Chitram, S.; Kumar, S.; Thenmalar, S. Enhancing Fire and Smoke Detection Using Deep Learning Techniques. *Eng. Proc.* **2024**, *62*, 7.
16. Ye, M.; Luo, Y. A deep convolution neural network fusing of color feature and spatio-temporal feature for smoke detection. *Multimed. Tools Appl.* **2024**, *83*, 22173–22187. [CrossRef]
17. Fang, L.; Zhang, Y.; Li, H.; Chen, B.; Sui, Y. Dynamic attention network for flame detection. *Electron. Lett.* **2023**, *59*, e12690. [CrossRef]
18. Kong, S.; Deng, J.; Yang, L.; Liu, Y. An attention-based dual-encoding network for fire flame detection using optical remote sensing. *Eng. Appl. Artif. Intell.* **2024**, *127*, 107238. [CrossRef]

19. Gu, K.; Xia, Z.; Qiao, J.; Lin, W. Deep dual-channel neural network for image-based smoke detection. *IEEE Trans. Multimed.* **2019**, *22*, 311–323. [CrossRef]
20. Khanam, R.; Hussain, M. Yolov11: An overview of the key architectural enhancements. *arXiv* **2024**, arXiv:2410.17725.
21. Wang, G.; Li, H.; Li, P.; Lang, X.; Feng, Y.; Ding, Z.; Xie, S. M^4SFWD: A Multi-Faceted synthetic dataset for remote sensing forest wildfires detection. *Expert Syst. Appl.* **2024**, *248*, 123489. [CrossRef]
22. Wei, H.; Liu, X.; Xu, S.; Dai, Z.; Dai, Y.; Xu, X. DWRSeg: Rethinking efficient acquisition of multi-scale contextual information for real-time semantic segmentation. *arXiv* **2022**, arXiv:2212.01173.
23. Du, B.; Zhang, L. Target detection based on a dynamic subspace. *Pattern Recognit.* **2014**, *47*, 344–358. [CrossRef]
24. Yu, Z.; Huang, H.; Chen, W.; Su, Y.; Liu, Y.; Wang, X. Yolo-facev2: A scale and occlusion aware face detector. *Pattern Recognit.* **2024**, *155*, 110714. [CrossRef]
25. Sheng, K.; Chen, P. An efficient mixed attention module. *IET Comput. Vis.* **2023**, *17*, 496–507. [CrossRef]
26. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 12993–13000.
27. Gevorgyan, Z. SIoU loss: More powerful learning for bounding box regression. *arXiv* **2022**, arXiv:2205.12740.
28. Tong, Z.; Chen, Y.; Xu, Z.; Yu, R. Wise-IoU: Bounding box regression loss with dynamic focusing mechanism. *arXiv* **2023**, arXiv:2301.10051.
29. Chino, D.Y.; Avalhais, L.P.; Rodrigues, J.F.; Traina, A.J. Bowfire: Detection of Fire in Still Images by Integrating Pixel Color and Texture Analysis. In Proceedings of the 2015 28th SIBGRAPI Conference on Graphics, Patterns and Images, Salvador, Brazil, 26–29 August 2015; IEEE: New York, NY, USA, 2015; pp. 95–102.
30. de Venâncio, P.V.A.B.; Lisboa, A.C.; Barbosa, A.V. An automatic fire detection system based on deep convolutional neural networks for low-power, resource-constrained devices. *Neural Comput. Appl.* **2022**, *34*, 15349–15368. [CrossRef]
31. Shamsoshoara, A.; Afghah, F.; Razi, A.; Zheng, L.; Fulé, P.Z.; Blasch, E. Aerial imagery pile burn detection using deep learning: The FLAME dataset. *Comput. Netw.* **2021**, *193*, 108001. [CrossRef]
32. Henderson, P.; Ferrari, V. End-to-end Training of Object Class Detectors for Mean Average Precision. In *Proceedings of the Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, 20–24 November 2016*; Revised Selected Papers, Part V 13; Springer International Publishing: Cham, Switzerland, 2017; pp. 198–213.
33. 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]
34. Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; Tian, Q. Centernet: Keypoint Triplets for Object Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6569–6578.
35. Wang, C.; He, W.; Nie, Y.; Guo, J.; Liu, C.; Wang, Y.; Han, K. Gold-YOLO: Efficient object detector via gather-and-distribute mechanism. *Adv. Neural Inf. Process. Syst.* **2024**, *36*, 51094–51112.
36. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475.
37. Kajita, E. Notes from the Field: Emergency Department Use During the Los Angeles County Wildfires, January 2025. *MMWR. Morb. Mortal. Wkly. Rep.* **2025**, *74*, 40–42. [CrossRef]
38. Seydi, S.T. Assessment of the January 2025 Los Angeles County wildfires: A multi-modal analysis of impact, response, and population exposure. *arXiv* **2025**, arXiv:2501.17880.

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.

GeoHazards

MDPI

Article

Predicting Railway Slope Failure Under Heavy Rainfall Using the Soil Moisture Extended Cohesive Damage Element Method

Sudath Loku-Pathirage [1], Jiye Chen [2,*] and Min Fu [3]

[1] Department of Earth Resource Engineering, University of Moratuwa, Moratuwa 10400, Sri Lanka
[2] School of Civil Engineering and Surveying, University of Portsmouth, Portsmouth PO1 3AH, UK
[3] College of Mechanical Engineering, Yanshan University, Qinhuangdao 066004, China
* Correspondence: jiye.chen@port.ac.uk

Abstract: Slope failure, as a natural disaster, can cause extensive human suffering and financial losses worldwide. This paper introduces a new soil moisture extended cohesive damage element (SMECDE) method to predict railway slope failure under heavy rainfall. A correlation between rainfall intensity and soil moisture content is first established to create an equivalence between the two. Considering slope failure mechanisms dominated by the loss of soil or the cohesion of slope materials due to heavy rainfall infiltration, the soil moisture decohesion model (SMDM) is developed using previous experimental data to express how soil cohesion varies with different soil moistures and depths. The SMDM is incorporated into the extended cohesive damage element (ECDE) method to fundamentally study slope failure mechanisms under varying soil moisture levels and depths. The proposed SMECDE approach is used to predict the failure propagation of a selected railway embankment slope at the critical soil moisture or rainfall intensity. This SMECDE failure prediction is validated using relevant data from previous fieldwork and meteorological reports on the critical rainfall intensity at the site. Additionally, the corresponding slope damage scale prediction is validated with a large plastic deformation analysis using the commercial FEM package ABAQUS.

Keywords: soil moisture—rainfall intensity correlation; soil moisture decohesion model; soil moisture extended cohesive damage element method; railway embankment slope failure

1. Introduction

Heavy rainfall is one of the major consequences of climate changes. Extensive environmental or infrastructural damage, as well as social decentralization, can occur due to continuous heavy rainfall. In the current scenario, slope failure is one of the major natural disasters. Slope failure includes both natural slope failures (landslides) and engineered slope failures, such as the collapse of railway embankments. Predicting slope failure under heavy rainfall is a significant challenge. Over the past few decades, many scientists have investigated slope failure under various conditions. The types of slope failure, remediation methods, and causes were discussed by Hamed Niroumand et al. (2012) [1]. Slope failure can be classified into four types of ground movement: slides (both translational and rotational), flows, falls, and topples. Slides and flows have the potential to cause extensive damage to infrastructure. Many factors affect slope stability, including climatic, geomorphological, geological, hydrogeological, and geotechnical factors, as well as site-specific stress and strain fields. Certain types of slope failure, such as translational and rotational movement, initiate when heavy rainfall causes a critical reduction in shear strength or

loss of cohesion at weak interfaces in layered terrain, leading to gravity-driven sliding or shear failure. In many countries worldwide, more than 90% of landslides occur during the summer due to heavy monsoons causing a loss of apparent cohesion (CGEIS, 2019 [2]; Froude et al., 2018 [3]). Much research has been undertaken on various elements of slope failure initiation, but current forecast models lack spatiotemporal accuracy in predicting slope failure. Geological models that forecast location, meteorological triggering conditions, and the potential extent of slope collapse have been elusive, and statistical models have limitations in identifying specific locations and the extent of impact for progressive slope failure (X Gu et al., 2023–2024, [4,5]). Making existing geotechnically based models more accurate requires geophysical data at a scale and certainty that is not generally available (Qiang, 2012) [6]. Previous research that measured deformation fields and creep using physical sensors to understand the evolution of strain and stress fields in potential slope failure regions and attempted to forecast slope failure occurrences using the critical deformation fields demonstrated the difficulty of forecasting slope failure propagation and extent (Teza et al., 2008) [7]. Conventional analytical or fracture mechanics models can only treat simple slope failures, such as translational landslides, and have difficulty forecasting progressive slope failure caused by critical rainfall (Milledge et al., 2014) [8]. Traditional finite element methods have limitations in forecasting triggering conditions, propagation, and the extent of slope failure due to implicit characterization of discontinuities (Donze et al., 2009) [9]. A standard cohesive zone model can only simulate the area of sliding land from a known failure surface, making it difficult to forecast landslide propagation from unknown failure surfaces (Quinn et al., 2012 [10]; Zhang et al., 2015 [11]). A particle finite element method (PFEM) (Zhang et al., 2019 [12]) has been developed to simulate retrogressive slope collapse using the elastoviscoplastic model with strain softening to simulate the progressive failure process and the extent of slope collapse, presented by large plastic deformation instead of sliding fracture propagation along an initially prepared failure surface. However, the PFEM has limitations in forecasting initiation and the starting location of slope failure under heavy rainfall conditions.

Slope failure in layered terrain or infrastructures can initiate and propagate anywhere along unknown failure surfaces. The hypothesis of the 'sliding failure mechanism' is that the critical reduction in shear strength or cohesion, along with the related shear fracture energy lost in soil at the interfaces due to rainfall infiltration in layered terrain, leads to slope failure. Rainfall infiltration directly causes changes in soil moisture, which is a critical factor in the reduction of soil cohesion. Soil moisture content can be defined as a volumetric soil moisture content, which represents the volume of water in the soil as a fraction or percentage of the total soil volume [13]. The soil moisture is mainly affected by rainfall intensity and soil depth. Other variables including rainfall duration and frequency, temperature, humidity, soil properties, land cover, water table depth, and underlying geology would also have influence. This is a rather complex topic. Progressively investigating different variables is a practical approach. This investigation focuses on the major factors, e.g., rainfall intensity and soil depth in a selected site with determined soil properties. The effect of short-intensive rainfall (average 2 h) on soil moisture and cohesion is considered by this study. The effects of other variables on soil moisture can be studied in detail in future work.

The correlation between rainfall intensity and moisture content was initially indicated by Kirsten (1997) [14], MPhi Thesis (2006) [15], and Andrzej et al. (2021) [16]. In this paper, a novel approach is developed by incorporating the proposed soil moisture decohesion model (SMDM) into the extended cohesive damage element (ECDE) method to predict progressive railway embankment slope failure under heavy rainfall. This prediction of slope failure under heavy rainfall fills a research gap that was not addressed by previous investigations.

The outcomes of this paper will be used by railway management sectors to make decisions based on weather forecasts. This paper first correlates rainfall intensity with soil moisture using weather forecast data. Then, the investigation of rainfall intensity effects shifts to the investigation of soil moisture effects on soil cohesion reduction. Considering that slope failure mechanisms involve the loss of soil or slope material cohesion due to heavy rainfall, the SMDM is developed to link soil cohesion with soil moisture and depth below the ground surface. Under rainfall conditions, soil moisture decreases with increasing soil depth, soil cohesion decreases with increasing soil moisture, and soil cohesion increases with increasing soil depth. The SMDM is incorporated into the ECDE method to forecast slope failure under heavy rainfall. As an example, the proposed SMECDE approach is applied to predict damage initiation, damage belt formation, and the damage scale of a selected railway embankment slope in the UK Kent region under varying soil moistures or rainfall intensities. The prediction shows that the critical rainfall intensity and soil moisture are 66 mm/h and 35%, respectively. These critical values are identified using previous field surveys and meteorological data. This paper also demonstrates that a soil moisture of 31% is an important transition point significantly changing the initial stiffness and damage scale of the investigated railway embankment slope. The corresponding slope damage scale prediction is validated with a large plastic deformation analysis using the commercial FEM package ABAQUS 2020.

2. The Correlation Between Rainfall Intensity and Soil Moisture Content

Usually, the weather-forecasted soil moisture is measured based on natural moisture content, α (%), in the ground surface or very shallow soil layer, which represents the percentage of water in the soil. Rainfall intensity, λ (mm/h), is measured using a standard rainfall measurement cylinder, based on the height of collected water per hour. Both α and λ are part of the weather-forecasted data. Figure 1 shows an example of time histories for rainfall intensity and soil moisture over a certain time period in the Kent region, UK, provided by the website of the European Centre for Medium-Range Weather Forecasts (ECMWF) [17].

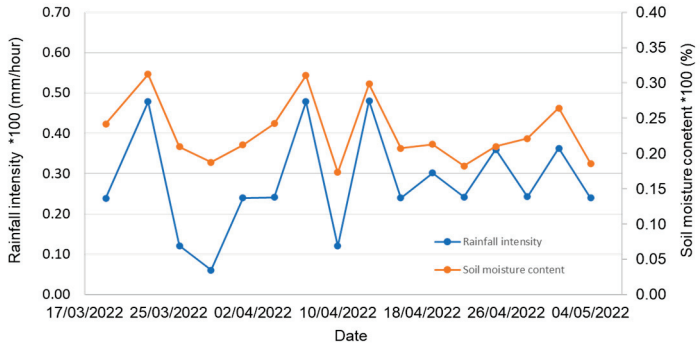


Figure 1. Rainfall intensity and soil moistures time histories.

This paper proposes a basic mathematic model as shown in Equation (1) to correlate soil moisture content with rainfall intensity using the meteorological data shown in Figure 1.

$$\alpha = k\lambda^n \quad (1)$$

Here, n and k are correlational parameters varying with locations and time periods. They could be taken as n = (0.1, 0.2, 0.3, 0.4, 0.5, ...) and k = (1, 1/2, 1/3, 1/4, 1/5, ...). The detailed meteorological data of time histories shown in Figure 1 are given in Table 1. When

n and k are taken as 0.21 and 1/3, respectively, using a curve fitting approach, Equation (1) becomes Equation (2).

$$\alpha = (1/3)\ \lambda^{0.21} \tag{2}$$

Table 1 also shows the calculated values of soil moisture α using Equation (2). The averaged error from Equation (2) is about 3% compared to forecasted values of soil moisture, which indicates that Equation (2) is a good correlation for rainfall intensity and soil moisture based on the provided meteorological data at Kent region, UK, from March to May 2022.

Table 1. Meteorological data from March to May 2022 at Kent region, UK.

Date	20/03	23/03	26/03	29/03	01/04	04/04	07/04	10/04	13/04	16/04	19/04	22/04	25/04	28/04	01/05	04/05
λ (mm/h)	0.2392	0.4785	0.1214	0.061	0.2396	0.2412	0.4785	0.1203	0.4805	0.2396	0.3021	0.2425	0.3596	0.2427	0.3631	0.2398
α (%)	0.2421	0.3121	0.2097	0.1875	0.2123	0.2423	0.3105	0.1734	0.2984	0.2072	0.2132	0.1823	0.2101	0.2214	0.2642	0.1857
Equation (2)	0.2475	0.2610	0.2401	0.2345	0.2407	0.2475	0.2607	0.2307	0.2586	0.2395	0.2409	0.2332	0.2402	0.2429	0.2520	0.2341

3. The Soil Moisture Decohesion Model (SMDM)

A correlational model is generally proposed as a polynomial function between soil moisture α and soil cohesion C in the ground surface as shown in Equation (3).

$$C = p\alpha^2 + q\alpha + s \tag{3}$$

Here, p, q, and s are correlational parameters. Combining Equations (1) and (3) results in Equation (4), which is a correlation between soil cohesion and rainfall intensity λ.

$$C = p(k\lambda^n)^2 + q(k\lambda^n) + s \tag{4}$$

Define $w = p \times k^2$, $t = q \times k^2$, and Equation (4) become Equation (5) as below.

$$C = w\lambda^{2n} + t\lambda^n + s \tag{5}$$

In Equation (3), correlational parameters p, q, and s can be determined with experimental data through testing soil samples with variable water contents in the laboratory. Equation (5) is an incurred correlation between soil cohesion and rainfall intensity. Its parameters w, t, and s can be determined indirectly using Equations (1) and (3) with assistance from the curve fitting approach used for determining parameters in Equation (2).

Considering the shallow soil layer underground, previous work [18] investigated how soil moisture α almost linearly decreases with increasing soil depth, *h* (meters), as shown in Figure 2. It can be seen from Figure 2 that α decreases from the ground surface moisture 21.654% along the linear path with a degradational ratio of −0.0417. It should be noted that the points shown on the figure are tested data from previous experiment work [18–21] (same for Figures 2–6). This was studied for native grassland and artificial grassland under rainfall conditions through a case study of a typical loess hilly area in the Anjiapo catchment of China during the 2010 growing season. Soil cohesion decreases with increasing soil moisture [18–20]. As an example, Figure 3 shows that soil cohesion decreases with increasing soil moisture at the ground surface. It is a polynomial decreasing function that is generally expressed using Equation (3). It should be noticed that this correlation between soil cohesion and moisture was conducted for the case of short-intensive rainfall (average 2 h) by previous experimental investigations [18–20]. Furthermore, soil cohesion increases with increasing soil depth [21] under rainfall conditions. A correlation for loamy soil with 25% moisture can be proposed based on previous test work [18–21], as shown in

Figure 4. Equation (6) mathematically expresses the linear correlation between soil cohesion and depth below the ground surface.

$$C_h = C + \Gamma h \tag{6}$$

Equation (3), together with Equation (6), is referred to as the general soil moisture decohesion model (SMDM). The parameters shown in Equation (3) vary with different soil types. Therefore, the parameters given in Equations (3)–(6) influence soil cohesion values for different soils at varying soil moisture levels and depths. Different types of soil can be identified by their liquid limit (LL). Previous experimental work [18] and Matsushi et al. (2006) [20] investigated the basic influence of soil moisture on cohesion variation. As an example, this paper selects a railway embankment slope in the UK Kent region for investigation, as Kent is one of the regions with the highest risk of railway slope failure, with over 4000 recorded slope failures in the past few decades. The soil type in the Kent region is primarily loamy soil with a liquid limit of 52%. The soil cohesion–moisture correlation in the Kent region is generated by interpolation using basic data from previous work [18,20], as shown in Figure 3. Thus, the correlation parameters p, q, and s in Equation (3) can be determined through curve fitting using the data shown in Figure 3, and their values are provided in Table 2. Using Equations (2) and (3), and with the assistance of the curve fitting approach, the correlation parameters w, t, and s in Equation (5) can also be determined, with their values provided in Table 3.

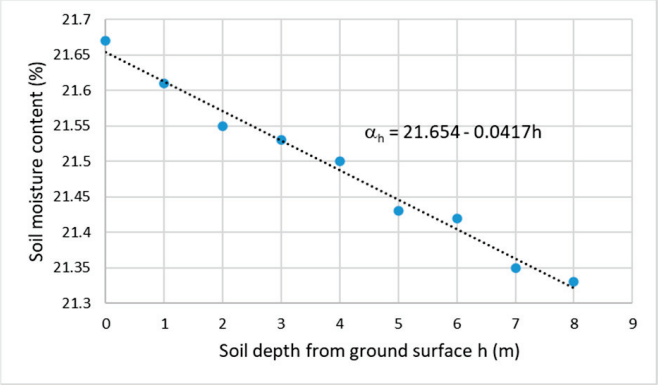


Figure 2. Soil moisture vs. depth underground [16].

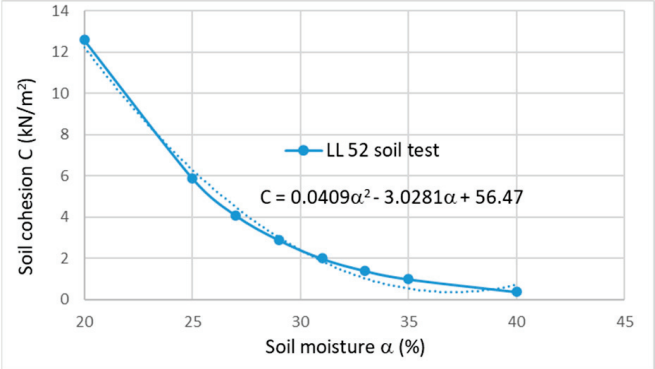


Figure 3. Soil cohesion vs. moisture for loamy soil.

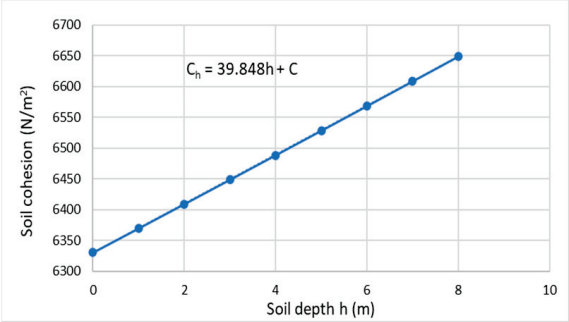


Figure 4. Loamy soil cohesion vs. depth at moisture 25%.

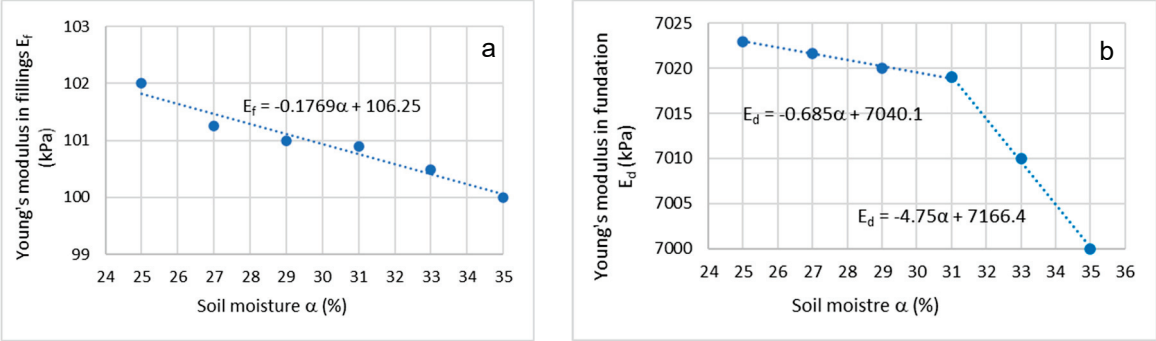


Figure 5. Young's modulus vs. soil moisture for embankment: (**a**) fillings, (**b**) foundation.

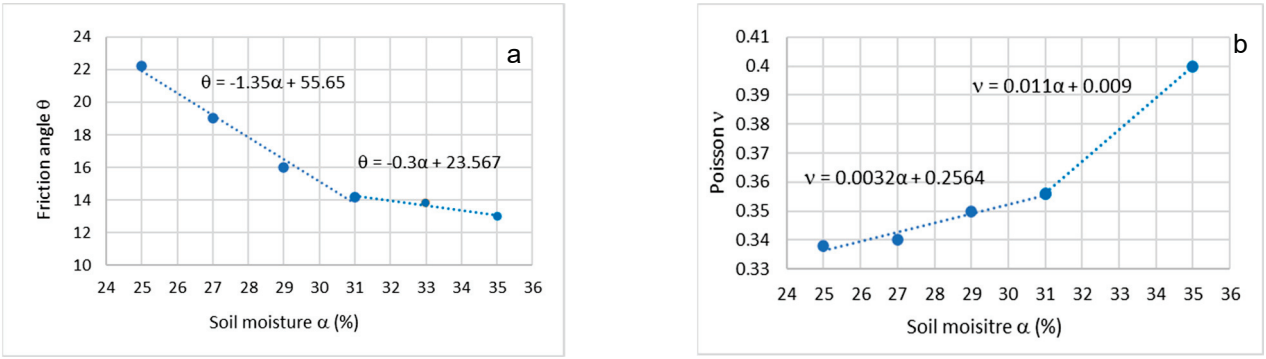


Figure 6. (**a**) Loamy soil's friction angle, (**b**) Poisson ratio vs. moisture content.

Table 2. The values for p, q, s and R^2.

Soil Type	p	q	s	R^2—Correlation Coefficient
LL52	0.0409	3.0281	56.47	0.9917

Table 3. The values for w, t, s and R^2.

Soil Type	w	t	s	R^2—Correlation Coefficient
LL52	0.004544	1.009367	56.47	0.9917

Using Equation (3) and parameters in in Table 2, Equation (3) can be formed with details for soil LL52 as shown in Equation (7).

$$C = 0.0409\alpha^2 - 3.0281\alpha + 56.47 \quad (7)$$

Equation (7) reflects the behavior of soil cohesion reduction when increasing moisture as a polynomial function shown in Figure 3 in the ground surface or very shallow soil layer. Using Equations (2) and (5) and Table 3, the incurred correlation between soil cohesion and rainfall intensity in Kent region can be expressed using Equation (8).

$$C = 0.004544\lambda^{0.6} + 1.009367\lambda^{0.3} + 56.4 \quad (8)$$

In Equation (6), C_h is the soil cohesion at the level that is h meters below ground surface, C is the cohesion in the ground surface, and Γ is a linear incredible parameter varying with different soil types. Here, Γ is taken as 0.004 based on the data analysis of the previous work [18–21]. Thus, a linear correlation between soil cohesion and depth using Equation (6) is shown in Figure 4 for soil LL52 in the Kent region. It should be noticed that Equation (6) is valid under rainfall conditions, and there is no consideration of underground water resources.

4. Soil Material Properties

Most soils with LL are commonly recognized as a porous material consisting of particles (or grains), which are held together by cohesion. In general, soil's LL, moisture content, particle scales, and density mainly determine the soil mechanical properties. This paper will focus on investigating the effects of soil moisture and depth on mechanical performance. The railway embankment slope consists of embankment filings and foundation, and the top layer consists of ash/ballast. It should be noted that embankment filings and foundation are the main body of the slope structure with manually prepared slope engineering materials. These materials would be local loamy soil mixed with others, including natural aggregates, thus named as loamy soil composites in this paper for a simple classification. In the UK railway embankment slope, aggregates with varying size from 5 mm to 40 mm are normally used to build the fillings, while the foundation layer consists of sand, gravel, or well-graded soil with particles up to 75 mm. More detailed information (e.g., grain size distribution, permeability, and compactness) of these engineering slope materials can be found in the references [18–21]. Their properties include soil moisture (α), soil types with different liquid limits (LL), cohesion (C), Young's modulus (E), friction angle (θ) and Poisson ratio (ν). The effects given their variable values with different moisture contents and soil depths on slope failure propagation of a selected railway embankment slope with geometrical slope angle, shape, and size are investigated in this paper. Using previous experimental data (Stirling 2014 [22], Tetsuya 2017 [23]), the Young's modulus (E_f) of embankment fillings varies with soil moisture (α) and can be plotted as shown in Figure 5a. It can be seen from Figure 5a that the Young's modulus (E_f) of embankment fillings of a loam soil type is gradually and almost linearly reduced when increasing soil moisture α, which can be expressed using a curve fitting equation as shown in Equation (9). Figure 5a shows a deceasing correlation from E_f = 106.25 (KPa) at 25% moisture down to E_f = 100 (KPa) with a ratio of −0.1769.

$$E_f = -0.1769\alpha + 106.25 \quad (9)$$

Using the same references [22,23], the Young's modulus (E_d) of the embankment foundation varying with soil moisture α is shown in Figure 5b. A bi-linear correlation between E_d and α can be estimated using Equation (10).

$$E_d = \begin{matrix} -0.685\alpha + 7040.1 \\ -4.75\alpha + 7166.4 \end{matrix} \Bigg\} \quad (10)$$

Figure 5b shows that E_d slowly decreases by a ratio of -0.685 from 7040.1 KPa at 25% moisture before it reaches a moisture of 31%. Then, it rapidly drops by a ratio of -4.75 to 7000 KPa at a moisture of 35%. Approximately 31% moisture would cause a significant change to the Young's modulus of the embankment foundation. This correlation aligns with previous work (Stirling, 2014; Tetsuya, 2017 [22,23]). Referring to the previous work by Li (2021) and Zhao (2018) [24,25], the friction angle θ of loamy soil varying with moisture content α is evaluated for the Kent region and shown in Figure 6a. This can also be estimated using a bi-linear function shown in Equation (11).

$$\theta = \left.\begin{matrix} -1.35\alpha + 55.65 \\ -0.3\alpha + 23.567 \end{matrix}\right\} \quad (11)$$

This soil friction angle θ is quickly reduced when the moisture content increases from 25% to 31%, followed by a slow reduction process. Here, it takes 12.5 degrees when α = 35%. In addition, 31% moisture is also an important point as two different degradation ratios, namely, -0.135 and -0.3, estimate the reduction speed before and after that moisture point. After that, the friction angle would become stable. Meanwhile, the correlation between loamy soil's Poisson ratio ν and moisture content α can be estimated in terms of previous work (Thota 2021 [26], Oh W. T. 2011 [27]), which is shown in Figure 6b. It can be seen from Figure 6b that Poisson's ratio of loamy soil increases as the moisture content increases. Its correlation can also be expressed roughly using another bi-linear function as shown in Equation (12).

$$\nu = \left.\begin{matrix} 0.0032\alpha + 0.2564 \\ 0.011\alpha + 0.009 \end{matrix}\right\} \quad (12)$$

It can be seen from Figure 6b that there is also important moisture point α (31%). Specifically, a slowly increased ratio of 0.0032 can be seen before 31%, followed by a significant increased ratio 0.011 after 31%. Poisson's ratio of loamy soil is the only parameter that increases when the moisture content increases, which implies that this soil could have more plastic strain in any direction. In general, increases in soil moisture result in decreased soil cohesion, Young's modulus, and friction angle. These changes will cause a reduction in the soil's stiffness and strength. Thus, soil gradually loses cohesion and stability and finally collapses under heavy rainfall and high moisture content. It can be seen from the above analysis that the Young's modulus of fillings are stably decreased as soil moisture increases. Here, 31% soil moisture is an important transition point with significant increases in the foundation's Young's modulus and friction angle and deceases in Poisson's ratio of the studied railway embankment slope.

It should be noted that, like soil cohesion, which varies with depth, basic soil properties, including Young's modulus, also vary with depth. Other material parameters, e.g., friction angle and Poisson's ratio, are less varied in depth considering the materials in the railway embankment slope because they are manually prepared. Therefore, two layers with different material properties, including cohesion and the Young's modulus for embankment fillings and foundation, are applied. Additionally, two groups of constant materials for the friction angle and Poisson's ratio, treated as depth-averaged materials, are proposed for the embankment fillings and foundation in this investigation. Alternatively, the correlation between soil moisture and depth obtained from Figure 2, $\alpha_h = \alpha - \phi h$ (ϕ reflects degradation of moisture from ground surface and taken as 0.04 in terms of references [18–21]), can be used together with Equations (9)–(12) to calculate different parameters varying with depth.

5. The Basic SMECDE Equations

Physically, slope sliding is a progressive failure process involving shear damage and fracture propagation. The Soil Moisture Extended Cohesive Damage Element (SMECDE) method integrates the computational Extended Cohesive Damage Element (ECDE) model with the Soil Moisture Distribution Model (SMDM) to analyze railway embankment slope failure propagation while accounting for soil moisture effects. This approach includes predictions for slope failure initiation, failure location, and the progressively formed failure area.

The ECDE method, developed by the corresponding author Chen and his colleagues, is designed for predicting damage and fracture propagation in engineering materials and structures (Chen et al., 2022; Li and Chen, 2016, 2017) [28–35]. The fundamental theory of the SMECDE is detailed below. Specifically, the incorporation of the SMDM is discussed in Section 3, and its applications is outlined in Section 7.

Considering that soil material properties vary with soil moisture, the soil element stiffness K and displacement field u are a function of soil moisture α. Before cracks occur at low soil moisture content, u is the ordinary displacement field. Here, K and u can be used to form a basic equilibrium equation with applied force F shown in Equation (13).

$$\mathrm{K\,u = F} \tag{13}$$

Here, F is also a function of soil moisture α due to the variable amount of water in the soil. When soil moisture reaches the critical content, crack/fracture will occur; thus, the displacement field with discontinuity must consider approximating the crack/fracture propagation in the slope body. Equation (14) shows the test and trial function of the discontinuous displacement field.

$$\mathrm{u(x)} = u^{s}(\mathrm{x}) + Step_{\Gamma_d}(x)\alpha(x) \tag{14}$$

In Equation (14), the ordinary displacement field and the displacement jump can be accounted for using $\mathrm{u^s(x)}$ and α(x), respectively. A step function $\mathrm{Step_{\Gamma d}(x)}$ is used to reflect a physical jump when the element is completely separated as a crack or strong discontinuity. Equation (15) shows a discrete FEM form to approximate the displacement field.

$$\mathrm{u^h(x)} = \sum_{\mathrm{i}\in\mathrm{I}} \mathrm{N_i(x)u_i} + \sum_{\mathrm{j}\in\mathrm{I}} \mathrm{N_j(x)(H_{\Gamma_d}(x) - H_{\Gamma_d}(x_i))a_i} \tag{15}$$

In Equation (15), $\mathrm{u_i}$ and $\mathrm{a_i}$ are ordinary displacement DoFs and additional enrichment DoFs, respectively, used to present the displacement jump or crack, and $\mathrm{N_i}$ and $\mathrm{N_j}$ are standard FEM shape functions. Equation (16) shows the basic Heaviside step function $\mathrm{H_{\Gamma_d}}$.

$$\mathrm{H_{\Gamma_d}(x)} = \begin{cases} 1, \ \forall \mathrm{x} \in \Omega^{+} \\ 0, \ \forall \mathrm{x} \in \Omega^{-} \end{cases} \tag{16}$$

Define $\mathrm{H_{\Gamma_d}(x)} - \mathrm{H_{\Gamma_d}}(x_i)$ as a shifted Heaviside function for modelling a crack, expressing the displacement jump as show in Equation (17).

$$\mathrm{\Delta u = a_1N_1(x)(H_{\Gamma_d}(x) - H_{\Gamma_d}(x_1)) + a_2N_2(x)(H_{\Gamma_d}(x) - H_{\Gamma_d}(x_2))} \tag{17}$$

Equation (17) shows a combination with two enriched items, which results in a real displacement with a jump Δu. A single value for any crack geometry at any node can be ensured by this shifted enrichment at the displacement basis. It should be notes that the enriched contribution has a certain value at material integration points but vanishes

at enriched nodes. Equation (13) can be changed to the discrete form of the equilibrium equation expressed by Equation (18) using the Bubnov–Galerkin method and the weak form of equilibrium equation for static analysis.

$$\begin{bmatrix} K^{uu} & K^{ua} \\ K^{au} & K^{aa} \end{bmatrix} \begin{bmatrix} u \\ a \end{bmatrix} = \begin{bmatrix} f^{u}_{ext} \\ f^{a}_{ext} \end{bmatrix} \tag{18}$$

In Equation (18), the stiffness matrices relating to the standard FE approximation and the enriched approximation can be accounted for using K^{uu} and K^{aa}, respectively. K^{ua} and K^{au} represent the coupling between the standard FE approximation and the enriched approximation; while, the equivalent nodal force vectors for standard FEM DoFs and enriched DoFs are approximated using f^{u}_{ext} and f^{a}_{ext}, respectively. Equation (19) gives the calculation of each submatrix required by Equation (18).

$$\begin{cases} K^{uu} = \int_{\Omega^h} B^T D B d\Omega \\ K^{ua} = \int_{\Omega^h} B^T D B^a d\Omega \\ K^{au} = \int_{\Omega^h} (B^a)^T D B d\Omega \\ K^{aa} = \int_{\Omega^h} (B^a)^T D B^a d\Omega \end{cases} \tag{19}$$

In Equation (19), details for the basic constitutive material matrices D, the deformation gradient matrices B produced for strain calculation, and the deformation gradient matrices B^a relating to the enriched DoFs can be obtained from [28–35]. It should be noted that their calculations use varied material properties at the critical soil moisture, as discussed in Sections 3 and 4. It should be noted that the simulated crack is assumed as a cohesive crack, and the corresponding discontinuous boundary is a cohesive crack boundary. Thus, using Equation (20), the equivalent nodal force vectors in Equation (18) can be calculated.

$$\begin{aligned} f^{u}_{ext} &= \int_{\Gamma_d} N^T \bar{t} d\Gamma \\ f^{a}_{ext} &= \int_{\Gamma_d} N^T \left(H_{\Gamma^{crack}_{coh}}(x,y) - H_{\Gamma^{crack}_{coh}}(x_i,y_i) \right) \bar{t} d\Gamma + f_{coh} \end{aligned} \tag{20}$$

The calculation of the internal nodal force vector due to cohesive traction t on the crack surface Γ_d, as shown in Figure 7, can be expressed using Equation (21).

$$\begin{aligned} f_{coh} = & \int_{\Gamma_{d^+}} (H_{\Gamma_d}(x,y) - H_{\Gamma_d}(x_1,y_1)) N^T t d\Gamma + \int_{\Gamma_{d^-}} (H_{\Gamma_d}(x,y) - H_{\Gamma_d}(x_1,y_1)) N^T (-t) d\Gamma \\ & = \int_{\Gamma_d} N^T t d\Gamma \end{aligned} \tag{21}$$

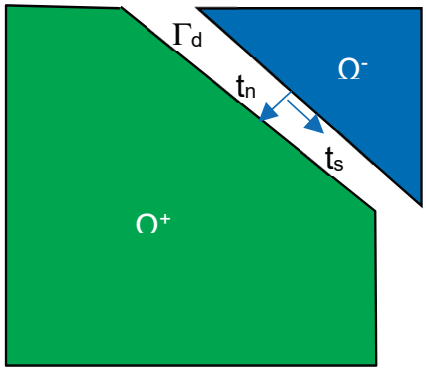


Figure 7. Demonstration of the cohesive crack surface Γ_d between Ω^+ and Ω^- in a potentially cracked body.

The calculation for the internal nodal force vector using the standard Gaussian integration scheme is presented in Equation (22).

$$f_{coh} = \int_{\Gamma_d} N^T t d\Gamma = \sum_{i=1}^{n} N^T(\xi_i) t(\xi_i) w(\xi_i) \tag{22}$$

Here, ξ_i is the coordinate of the Gauss integration point i ($1 \leq i \leq n$), and $w(\xi_i)$ is a weight function. Using Equations (20)–(22), Equation (18) can be changed to Equation (23).

$$\begin{bmatrix} K^{uu} & K^{ua} \\ K^{au} & K^{aa} \end{bmatrix} \begin{bmatrix} u \\ a \end{bmatrix} = \begin{bmatrix} f_{ext}^{u} \\ \int_{\Gamma_d} N^T \left(H_{\Gamma_{coh}^{crack}}(x, y) - H_{\Gamma_{coh}^{crack}}(x_i, y_i) \right) \bar{t} d\Gamma \end{bmatrix} + \begin{bmatrix} 0 \\ f_{coh} \end{bmatrix} \tag{23}$$

A mathematic matrix operation is used in this work to eliminate the additional enrichment a to form a fully condensed equilibrium system and to improve computing efficiency especially at the structural level. Thus, Equation (24) is obtained to present equilibrium equation using the standard FEM unknown quantities.

$$(K^{uu} - K^{ua}(K^{aa})^{-1}K^{au})u = f_{ext}^{u} - K^{ua}(K^{aa})^{-1}f_{coh} \tag{24}$$

It can be seen from Equation (24) that the discontinuity-related enriched DoFs a make negative contributions to element stiffness on the left side and force on the right side. This change is obviously due to materials cracking at the critical soil moisture. Assume an equivalent elemental stiffness matrix K^e and equivalent force vector F^e to replace corresponding items on the left and right side, as shown in Equations (25) and (26), respectively.

$$K^e = K^{uu} - K^{ua}(K^{aa})^{-1}K^{au} \tag{25}$$

$$F^e = f_{ext}^{u} - K^{ua}(K^{aa})^{-1}f_{coh} \tag{26}$$

Thus, the equivalent element equilibrium equation at the cracking stage can be simply expressed using Equation (27).

$$F^e = K^e u \tag{27}$$

It should be noted that Equation (23) assumes that no external load is applied to the cohesive crack. The SMECDE formulation enables the calculation of nodal displacements for cracked elements using standard FEM degrees of freedom (DoFs) only, thereby improving computational efficiency and convergence in nonlinear iterations at the structural level.

This proposed mathematical procedure is rigorous and effectively captures the entire damage evolution process, transitioning from weak to strong discontinuities. By eliminating the need for enriched DoFs, the SMECDE method can be seamlessly integrated into the standard commercial FEM software package ABAQUS using the user-defined element (UEL) feature. This integration involves incorporating the formulations of Equations (3), (6), and (7) along with Equations (13)–(27).

The developed SMECDE approach is applied to predict railway slope failure propagation at the selected site. Figure 8 illustrates the ECDE modeling process implemented in ABAQUS. The procedure consists of the following five steps:

a. Create the geometry and standard FEM meshes in an ABAQUS input (INP) file.
b. Apply ECDE elements within the standard FEM meshes using double meshing technology.
c. Generate a new INP data file with user-defined ECDE elements.

d. Execute the ECDE modeling using the ABAQUS nonlinear solver, such as the Newton–Raphson method with a line search algorithm.
e. Visualize failure outcomes by transferring results into a mesh representation based on strain contours.

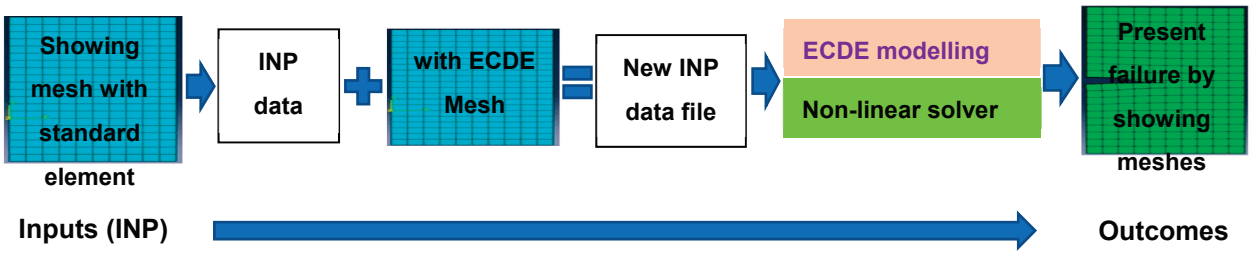


Figure 8. ECDE modelling process using ABAQUS.

6. Slope Failure Modes and Damage Criteria

Slope collapses are characterized by the downward sliding of slope materials along a weak surface. Figure 9a schematically illustrates a weak (slip) surface represented by red dashed lines (Mario Bačić et al., 2022 [36]). A real example of a railway embankment slope collapse in the Kent region of the UK is shown in Figure 9b. Generally, elements of slope collapse include the following features: a fracture at the top of the failed zone, flank fractures on the sides of the failed zone, and a dominant failure mode at the weak surface involving shear fractures and compressive damage. The fracture modes in slope collapse primarily represent mixed-mode damage. Specifically, the top fracture is a tensile (mode I) crack, the flank fractures are shear (mode II) cracks, and the failure at the weak surface exhibits mixed-mode damage, combining shear cracking and compressive crushing.

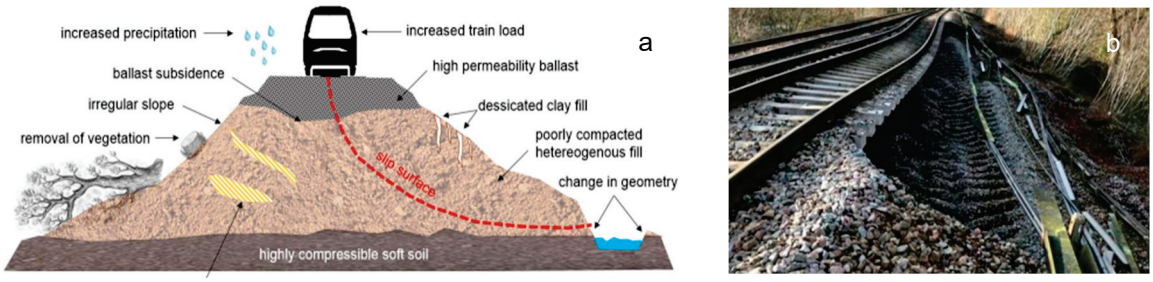


Figure 9. Railway embankment slope with a weak surface (**a**) and a case of failure in Kent, UK (**b**).

Based on observations, it is proposed that crack propagation at the weak surface is a form of mixed-mode damage in which porous materials are poorly connected (Gaume et al.,2018 [37]; Philipp et al., 2020 [38,39]). Before a slope collapse is released, certain conditions must be met, including the initiation and propagation of top tensile fractures, shear flank fractures, and mixed-mode damage in the weak layer. Once fractures or damage initiate, the slope materials above the weak layer are released if external compression and the self-weight-related driving force overcome the friction of the bed surface. Thus, slope failure propagation can be divided into two stages: damage initiation and crack propagation.

Referring to previous work (Chen, et al. 2022) [28], a general hybrid stress-based criterion for the mixed-mode damage initiation of porous materials is proposed as shown in Equation (28).

$$\frac{\sigma_n^2}{\sigma_{nc}^2} + \frac{\sigma_t^2}{\sigma_{tc}^2} + \frac{\sigma_c^2}{\sigma_{cc}^2} = 1 \quad (28)$$

Here, σ_n, σ_t, and σ_c are tensile, shear, and compressive stress, respectively. σ_{nc}, σ_{tc}, and σ_{cc} are their critical values or strengths. Considering soil material features and failure mechanisms, tensile stress can be ignored when evaluating damage initiation, as the loss of cohesion and the influence of compressive stress due to self-weight and external compressive loads dominate. Compressive stresses bring soil particles together and increase shear strength and shear fracture energy according to the basic soil mechanics shown in Equation (29), which presents a relationship among soil shear strength, cohesion, the friction angle, and normal stress.

$$\tau_f = C + \sigma_c \tan\phi \tag{29}$$

In Equation (29), normal compressive stress σ_c varies according to the position in embankment slope, and shear strength τ_f is the same as the σ_{tc} in Equation (28). In addition, cohesion C and friction angle ϕ vary with different soil types, moisture contents, and depths as noted in Sections 3 and 4. Since the effect of compressive stress can be translated into the enhanced shear strength using the basic soil mechanics, Equation (28) can be simplified as $\sigma_t = \tau_f$, and it will be used in the assessment of damage initiation in the railway embankment slope in the Kent region. In the second stage of crack propagation, the energy-based criteria are used to judge the mixed-mode crack propagation, as shown in Equation (30).

$$\frac{G_n^2}{G_{nc}^2} + \frac{G_t^2}{G_{tc}^2} + \frac{G_c^2}{G_{cc}^2} = 1 \tag{30}$$

Here, G_n, G_t, and G_c are strain energy release rates for tensile, shear, and compressive cases. G_{nc} and G_{tc} are corresponding critical values, i.e., fracture toughness for tensile and shear cases. G_{nc} and G_{tc} are used to assess the tensile and shear dominated fracture propagation on top and both sides of slab, respectively. In these two individual cases, Equation (30) is simplified as $G_n = G_{nc}$ and $G_t = G_{tc}$, respectively. G_{cc} is defined as a crushing toughness to assess compression damage propagation, and G_{tc}, together with G_{cc}, is used to judge mixed-mode damage propagation with coupling effects from shear crack and crushing damage. As mentioned in the previous section, compressive effects can be reflected by increased shear strength and shear fracture energy. Thus, Equation (30) can be simplified as $G_t = G_{tc}$ for judging the dominated shear crack propagation in the slope. This paper focus on the first stage for damage initiation and progressive formation of the failure area. In this first failure stage, the slope self-weight and external action-related driving force causes the internal stress state to reach the stress-based damage criteria.

7. Modelling Prediction

As an example, the railway embankment slope in the Kent region is selected for investigation in this paper. Figure 10 shows a half-configuration of the railway embankment slope, including its geometrical size, shape, and external load, utilized for the investigation due to its symmetrical features [18]. The material properties considered in this investigation include soil moisture, cohesion, Young's modulus, friction angle, and Poisson's ratio. These properties are treated as functions of soil moisture, following the equations and figures presented in Sections 3 and 5. Their detailed values are provided in Table 4. Figure 11 depicts a basic SMECDE model of half of the railway embankment slope, with the following boundary conditions: symmetrical restraints applied at the left edge and fixed restraints applied at the bottom of the model. The mesh density, as shown in Figure 11, varies with element sizes ranging from 0.1 to 0.25 m for good accuracy. The proposed SMECDE element, consisting of four nodes, is used in the numerical mesh for the simulation. This element has very good iteration accuracy for a relatively wide range of mess densities. Its detailed

mess independence can be seen in [31–35]. The modeling analysis includes both geostatic and large deformation analyses driven by gravity and train load. The railway train load is simplified as a linearly distributed pressure of 37.5 kN/m. Then, the disturbance from the train to the embankment will be overlooked.

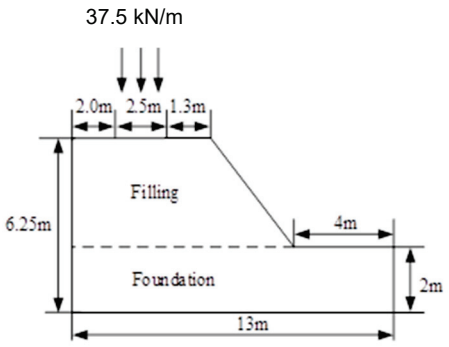


Figure 10. Half of the railway embankment slope.

Table 4. Material parameters varying with moisture.

Moisture α (%)	Cohesion C (N/m^2)	Young's Modulus E_f of the Embankment Fillings (N/m^2)	Young's Modulus E_d of the Embankment Foundation (N/m^2)	Soil Friction Angle θ (°)	Poisson's Ratio v
25	6330	101,828	7,022,975	22	0.34
27	4527	101,474	7,021,605	19	0.34
29	3052	101,120	7,020,235	17	0.35
31	1904	100,766	7,018,865	14	0.36
33	1083	100,412	7,009,650	14	0.37
35	589	100,059	7,000,150	13	0.39

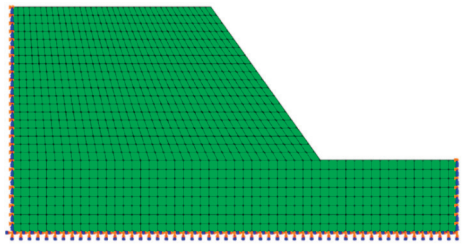


Figure 11. A model of the railway embankment slope.

There are six cases with varied soil moisture analyzed using the SMECDE approach. The ground surface soil moisture α varies: 25%, 27%, 29%, 31%, 33%, and 35%. The other corresponding material parameters are presented in Table 4. It should be noted that soil cohesion and the Young's modulus vary with soil depth at different moistures and can be calculated using the equations and figures presented in Sections 3 and 5. As an example, Table 5 shows their values at ground surface moisture of 35% varying with soil depths from 0 to 6 m (the railway embankment slope depth). In Table 5, E_f varies from 0 to 4 m for embankment fillings, while E_d varies from 5 to 6 m for the foundation. Predictions of damage initiation and the initial damage surface in the Kent railway embankment slope are presented using plastic shear strain contours.

Figure 12 shows predictions of variable damage initiation and the damage surface formed when increasing soil moisture α from 25% to 35%. At 25% moisture, there is only one point of damage at the corner between the embankment fillings and the basement. At 35% moisture, the initial damage finally reaches the point where the train pressure is

applied on the top of the slope, thus a significant damaged narrow belt is formed that is the so called initial damage surface. The initial damage surface separates the part of slope to the right above the initial failed surface from the rest of embankment body, but this should be identified using shear crack propagation in the second stage. The material parameters given in Table 5 at a soil moisture of 35% are used in the SMECDE modelling prediction.

Table 5. Soil cohesion and Young's modulus vary with depth with a ground surface moisture of 35%.

Depth h (m)	Moisture α (%)	Cohesion C (N/m^2)	Young's Modulus E_f from 0 to 4 m in the Fillings (N/m^2)	Young's Modulus E_d from 5 to 6 m in the Foundation (N/m^2)	Soil Friction Angle θ (°)	Poisson's Ratio v
0	35.00	589	100,059		13.07	0.39
1	34.96	596	100,066		13.08	0.39
2	34.92	602	100,073		13.09	0.39
3	34.88	609	100,080		13.1	0.39
4	34.84	616	100,087		13.11	0.39
5	34.80	624		7,001,100	13.12	0.39
6	34.76	631		7,001,290	13.14	0.39

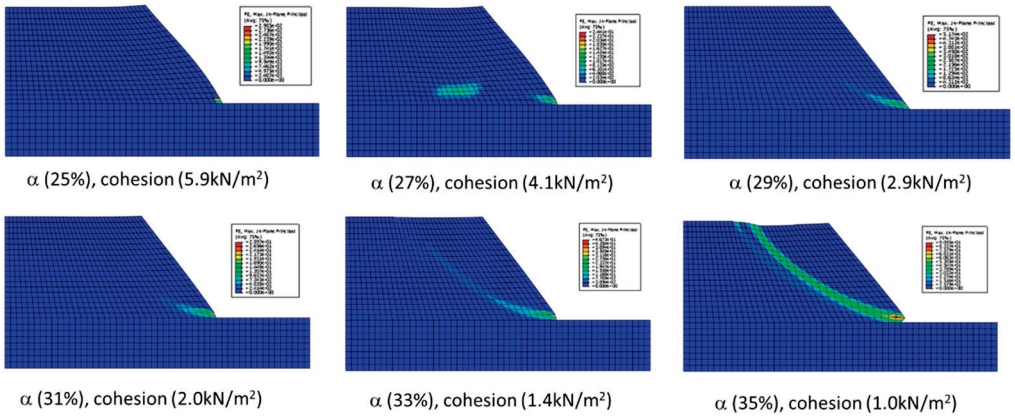


Figure 12. Predicted variable of initial damage surfaces at different soil moistures.

8. Damage Scale of the Railway Embankment Slope and Validation of Modelling Prediction

Using the outcomes from the SMECDE modelling analysis in Section 7, the damage scale β can be estimated using the initial damaged belt area presented in Figure 13 at varied soil moistures or rainfall intensities. The damage scale β shown in Figure 13 increases from 0 up to 1, which means damage starts at 25% moisture, and total damages occur at 35% moisture, which is identified by the SMECDE failure prediction (Figure 12). Figure 13a shows the damage scale β of the Kent railway embankment slope varying with soil moisture α (%). It can be seen from Figure 13a that the damage scale β stably increases from 25% to 31% moisture, then rapidly increases from 31% to 33% moisture, and finally reaches 1 at 35% moisture, which is identified as the failure state. Figure 13b shows the damage scale β varying with rainfall intensity λ (mm/h). Obviously, the damage scale β almost linearly increases between rainfall intensities of 15 and 40 (mm/h) and significantly increases after 40 (mm/h) rainfall intensity. This is followed by stably increasing to 1 at a rainfall intensity of 66 (mm/h). This rainfall intensity was considered as a critical rainfall value in previous reports [35]. Using Equation (1) with correlative parameters n = 0.23 and k = 0.385, a correlation between rainfall intensity and soil moisture, shown in Figure 13a,b, can be established using Equation (31).

To validate the failure prediction model of the Kent railway embankment slope, there have been no detailed reports investigating the slope failure using either laboratory or field work to date. However, previous meteorological data collected by [37–39] showed that the rainfall intensity varied from 60 to 70 mm/h when the railway embankment slope failed at Kent region, as shown in Figure 9b. This would validate the damage scale prediction in this investigation with a critical rainfall intensity of 66 (mm/h).

$$\lambda = \left(\frac{\alpha}{0.358}\right)^{\frac{1}{0.23}} \tag{31}$$

Furthermore, the standard commercial FEM package ABAQUS is used to partly validate the SMECDE failure prediction of the Kent railway embankment slope presented in Section 7. Based on the same numerical mesh (Figure 10), a standard plain strain element of CPE4I in ABAQUS and the material parameters given in Tables 4 and 5 are employed to model the large deformation under the same loading and restrained conditions that are used in the SMECDE modelling for initial damage analysis. Figure 14 shows the outcomes of displacement contours of the studied slope at the varied soil moistures ranging from 25% to 35%. At 35% soil moisture, the displacement contour (the last one in Figure 14) presents a large deformation at the right top body, which also forms a failure surface between the right top body and the rest of the slope. This identifies the SMECDE prediction of the initial damage surface where there is a potential shear crack propagation. However, the standard plain strain element in ABAQUS cannot simulate the shear crack propagation along the failed surface or explicitly predict the failure load to cause the shear crack propagation in slopes.

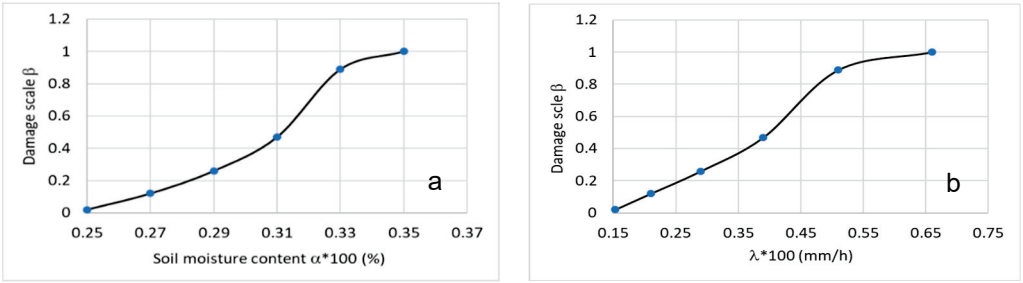


Figure 13. Predicted damage scale varies with (**a**) soil moisture, (**b**) rainfall intensity.

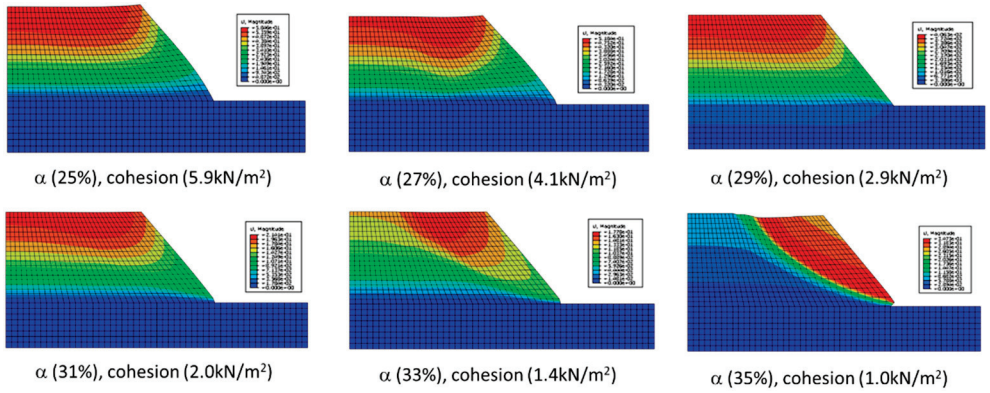


Figure 14. Predicted large deformations at different soil moistures.

9. Conclusions and Future Work

The developed SMECDE method was successfully applied to predict damage initiation and the progressive formation of the failure surface at the selected Kent railway embankment slope using meteorological data related to soil moisture and rainfall. This

investigation examined the embankment slope failure mechanisms in the Kent region under varying soil moisture levels and rainfall intensities.

For the first time, this study associated initial damage and the potential failure surface of a slope with forecasted rainfall intensity and soil moisture. It also explored the scale of slope damage as a function of soil moisture or rainfall intensity. The findings revealed that in the Kent region, when soil moisture reached 35%, the embankment slope was at a high risk of failure, with a corresponding critical rainfall intensity estimated at 66 mm/hour. The developed SMECDE method offers a promising tool for issuing railway slope failure warnings based on weather forecasts. The SMDM was developed using data from previous investigations that considered national soil moisture content. Future research will focus on investigating the effects of time-dependent changes in soil moisture or rainfall intensity on soil cohesion reduction. Furthermore, a temporal and spatial digital network could be developed using the SMECDE approach. This network could consider various factors, such as soil types, slope angles, and sizes, to predict railway slope failures based on weather forecasts, thereby aiding decision-making processes.

Author Contributions: Conceptualization, J.C.; methodology, J.C. and S.L.-P.; software, S.L.-P. and J.C.; validation, J.C.; formal analysis, S.L.-P. and M.F.; investigation, S.L.-P., J.C. and M.F.; resources, J.C.; data curation, S.L.-P. and M.F.; writing—original draft preparation, J.C.; writing—review and editing, J.C.; visualization, J.C. and S.L.-P.; supervision, J.C.; project administration, J.C.; funding acquisition, not available. 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 this study are included in the article. Further inquiries can be directed to the corresponding author.

Acknowledgments: The authors appreciate Simon Abbott at NetworkRail UK for kindly providing data on the railway embankment slope used in this investigation.

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

References

1. Niroumand, H.; Kassim, K.A.; Ghafooripour, A.; Nazir, R.; Far, S.Y. Investigation of Slope Failures in Soil Mechanics. *Electron. J. Geotech. Eng.* **2012**, *17*, 2703–2718.
2. China Geological and Environment Information Site (CGEIS). 2019. Available online: http://www.cigem.cgs.gov.cn/was5/web/search?page=40&channelid=281063&orderby=RELEVANCE&perpage=10&outlinepage=10&andsen=&total=&orsen=&exclude=&searchscope=&timescope=&timescopecolumn=&idStartDate2=&idEndDate2=&searchDate=&orderby=RELEVANCE (accessed on 12 March 2024).
3. Froude, M.J.; Petley, D.N. Global fatal landslide occurrence from 2004 to 2016. *Nat. Hazards Earth Syst. Sci.* **2018**, *18*, 2161–2181. [CrossRef]
4. Xu, Q.; Shang, Y.; van Asch, T.; Wang, S.; Zhang, Z.; Dong, X. Observations from the large, rapid Yigong rock slide—Debris avalanche, Southeast Tibet. *Can. Geotech. J.* **2012**, *49*, 589–606. [CrossRef]
5. Gu, X.; Zhang, W.; Ou, Q.; Zhu, X.; Qin, C. Conditional random field-based stochastic analysis of unsaturated slope stability combining Hoffman method and Bayesian updating. *Eng. Geol.* **2024**, *330*, 107415. [CrossRef]
6. Gu, X.; Wang, L.; Ou, Q.; Zhang, W. Efficient stochastic analysis of unsaturated slopes subjected to various rainfall intensities and patterns. *Geosci. Front.* **2023**, *14*, 101490. [CrossRef]
7. Teza, G.; Pesci, A.; Genevois, R.; Galgaro, A. Characterization of landslide ground surface kinematics from terrestrial laser scanning and strain field computation. *Geomorphology* **2008**, *97*, 424–437. [CrossRef]
8. Milledge, D.G.; Bellugi, D.; McKean, J.A.; Densmore, A.L.; Dietrich, W.E. A multidimensional stability model for predicting shallow landslide size and shape across landscapes. *J. Geophys. Res. Earth Surf.* **2014**, *119*, 2481–2504. [CrossRef]
9. Donzé, F.V.; Richefeu, V.; Magnier, S.A. Advances in Discrete Element Method Applied to Soil, Rock and Concrete Mechanics. *Electron. J. Geotech. Eng.* **2009**, *8*, 44.
10. Quinn, P.E.; Diederichs, M.S.; Rowe, R.K.; Hutchinson, D.J. A new model for large landslides in sensitive clay using a fracture mechanics approach. *Can. Geotech. J.* **2011**, *48*, 1151–1162. [CrossRef]

11. Zhang, W.; Wang, D.; Randolph, M.F.; Puzrin, A.M. Catastrophic failure in planar landslides with a fully softened weak zone. *Geotechnique* **2015**, *65*, 755–769. [CrossRef]
12. Zhang, X.; Wang, L.; Krabbenhoft, K.; Tinti, S. A case study and implication: Particle finite element modelling of the 2010 Saint-Jude sensitive clay landslide. *Landslides* **2019**, *17*, 1117–1127. [CrossRef]
13. Reeb, J.E.; Milota, M.R. *Moisture Content by the Oven-Dry Method for Industrial Testing*; WDKA: Rotterdam, The Netherlands, 1999; Available online: https://ir.library.oregonstate.edu/concern/conference_proceedings_or_journals/fq977v782 (accessed on 12 March 2024).
14. Findell, K.L. An Analysis of the Relationships Between Soil Moisture, Rainfall, and Boundary Layer Conditions. Ph.D. Thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, 1997.
15. Albataineh, N. Slope Stability Analysis Using 2D and 3D Methods. Ph.D. Thesis, University of Akron, Akron, OH, USA, 2006.
16. Brandyk, A.; Szeląg, B.; Kiczko, A.; Krukowski, M.; Kozioł, A.; Piotrowski, J.; Majewski, G. In Search of a Soil Moisture Content Simulation Model: Mechanistic and Data Mining Approach Based on TDR Method Results. *Sensors* **2021**, *21*, 6819. [CrossRef]
17. European Centre for Medium-Range Weather Forecasts (ECMWF). Available online: https://www.ecmwf.int/en/forecasts (accessed on 12 March 2023).
18. Yu, Y.; Wei, W.; Chen, L.D.; Jia, F.Y.; Yang, L.; Zhang, H.D.; Feng, T.J. Responses of vertical soil moisture to rainfall pulses and land uses in a typical Loess Hilly area, China. *Solid Earth* **2015**, *6*, 595–608. [CrossRef]
19. Earthworks Asset Policy Development—Global Stability and Resilience Appraisal Report, Network Rail August 2017, Mott MacDonald 1999. Available online: https://www.geolsoc.org.uk/~/media/shared/documents/events/Past%20Meeting%20Resources/Ground%20Related%20Risk/Mellor%20Richard.pdf (accessed on 12 March 2024).
20. Matsushi, Y.; Matsukura, Y. Cohesion of unsaturated residual soils as a function of volumetric water content. *Bull. Eng. Geol. Environ.* **2006**, *65*, 449–455. [CrossRef]
21. Dong, Y.; Wu, Y.; Yin, J.; Wang, Y.; Gou, S. Investigation of Soil Shear-Strength Parameters and Prediction of the Collapse of Gully Walls in the Black Soil Region of Northeastern China. *Phys. Geogr.* **2012**, *32*, 161–178. [CrossRef]
22. Stirling, R.A. Multiphase Modelling of Desiccation Cracking in Compacted Soil. Ph.D. Thesis, Newcastle University, Newcastle upon Tyne, UK, 2014.
23. Tetsuya, T. Climatic effects on the stability of unsaturated soil slope in cold region. In Proceedings of the 19th International Conference on Soil Mechanics and Geotechnical Engineering, Seoul, Republic of Korea, 17–21 September 2017.
24. Li, J. Experimental Study of Triaxial Test of Unsaturated Expansive Soil Shear Strength. *IOP Conf. Ser. Earth Environ. Sci.* **2021**, *692*, 042010. [CrossRef]
25. Zhao, Y.; Duan, X.; Han, J.; Yang, K.; Xue, Y. The main influencing factors of soil mechanical characteristics of the Gravity erosion environment in the dry-hot valley of Jinsha river. *Open Chem.* **2018**, *16*, 796–809. [CrossRef]
26. Thota, S.K.; Cao, T.D.; Vahedifard, F. Poisson's Ratio Characteristic Curve of Unsaturated Soils. *J. Geotech. Geoenviron. Eng.* **2021**, *147*, 04020149. [CrossRef]
27. Oh, W.T.; Vanapalli, S.K. Relationship between Poisson's ratio and soil suction for unsaturated soils. In *Theory and Practice*; Jotisankasa, A., Sawangsuriya, A., Soralump, S., Mairaing, W., Eds.; Kasetsart University: Bangkok, Thailand, 2011; ISBN 978-616-7522-77-7.
28. Chen, J.; Fyffe, B.; Han, D.; Yang, S. Predicting mixed mode damage propagation in snowpack using the Extended cohesive damage element method. *Theor. Appl. Fract. Mech.* **2022**, *122*, 103567. [CrossRef]
29. Ghimire, S.; Chen, J. Predicting fracture mechanisms in synthetic foam sandwiches with multi-layered cores using extended cohesive damage model. *Eng. Fract. Mech.* **2020**, *223*, 106719. [CrossRef]
30. Ghimire, S.; Chen, J. An extended cohesive damage model study of geometrical ratio effects on failure mechanisms of functionally graded sandwiches with multi-layered cores. *Compos. Struct.* **2019**, *224*, 110999. [CrossRef]
31. Li, X.; Chen, J. An extended cohesive damage model for simulating arbitrary damage propagation in engineering materials. *Comput. J. Methods Appl. Mech. Eng.* **2017**, *315*, 744–759. [CrossRef]
32. Li, X.; Chen, J. An extended cohesive damage model for simulating multicrack propagation in fibre composites. *Compos. Struct.* **2016**, *143*, 1–8. [CrossRef]
33. Li, X.; Chen, J. The implementation of the extended cohesive damage model for Multicrack evolution in laminated Composites. *Compos. Struct.* **2016**, *139*, 68–76. [CrossRef]
34. Li, X.; Chen, J. A highly efficient prediction of delamination migration in laminated Composites using the extended cohesive damage model. *Compos. Struct.* **2017**, *160*, 712–721. [CrossRef]
35. Li, X.; Chen, J. A highly efficient numerical approach: Extended cohesive damage model for predicting multicrack propagation. *Mater. High Temp.* **2017**, *34*, 371–385. [CrossRef]
36. Bačić, M.; Juzbašić, A. Challenges of Railway Embankments Remediation. In Proceedings of the 5th International Conference on Civil Structural and Transportation Engineering, Virtual, 12–14 November 2020. [CrossRef]
37. Gaume, J.; Gast, T.; Teran, J.; van Herwijnen, A.; Jiang, C. Dynamic anticrack propagation in snow. *Nat. Commun.* **2018**, *9*, 3047. [CrossRef]

38. Rosendahl, P.L.; Weißgraeber, P. Modeling snow slab avalanches caused by weak-layer failure—Part 2: Coupled mixed-mode criterion for skier-triggered anticracks. *Cryosphere* **2020**, *14*, 131–145. [CrossRef]
39. Decision pursuant to Article 10(5) of Protocol 36 to The Treaty on the Functioning of the European Union. Available online: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/326699/41670_Cm_8897_Print_Ready.pdf (accessed on 12 March 2023).

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

Development and Application of Self-Supervised Machine Learning for Smoke Plume and Active Fire Identification from the Fire Influence on Regional to Global Environments and Air Quality Datasets

Nicholas LaHaye [1,2,*], Anastasija Easley [2,3], Kyongsik Yun [2], Hugo Lee [2], Erik Linstead [4,5], Michael J. Garay [2] and Olga V. Kalashnikova [2]

[1] Spatial Informatics Group, LLC., Pleasanton, CA 94566, USA
[2] Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA 91101, USA; kyongsik.yun@jpl.nasa.gov (K.Y.); huikyo.lee@jpl.nasa.gov (H.L.); michael.j.garay@gmail.com (M.J.G.); olga.kalashnikova@jpl.nasa.gov (O.V.K.)
[3] Department of Mathematics, College of Letters & Science, University of California at Berkeley, Berkeley, CA 94720, USA
[4] Fowler School of Engineering, Chapman University, Orange, CA 92866, USA; linstead@chapman.edu
[5] Machine Learning and Assistive Technology Laboratory (MLAT), Chapman University, Orange, CA 92866, USA
* Correspondence: nlahaye@sig-gis.com

Abstract: Fire Influence on Regional to Global Environments and Air Quality (FIREX-AQ) was a field campaign aimed at better understanding the impact of wildfires and agricultural fires on air quality and climate. The FIREX-AQ campaign took place in August 2019 and involved two aircraft and multiple coordinated satellite observations. This study applied and evaluated a self-supervised machine learning (ML) method for the active fire and smoke plume identification and tracking in the satellite and sub-orbital remote sensing datasets collected during the campaign. Our unique methodology combines remote sensing observations with different spatial and spectral resolutions. With as much as a 10% increase in agreement between our produced masks and high-certainty hand-labeled pixels, relative to evaluated operational products, the demonstrated approach successfully differentiates active fire pixels and smoke plumes from background imagery. This enables the generation of a per-instrument smoke and active fire mask product, as well as smoke and fire masks created from the fusion of selected data from independent instruments. This ML approach has the potential to enhance operational active wildfire monitoring systems and improve decision-making in air quality management through fast smoke plume identification and tracking and could improve climate impact studies through fusion data from independent instruments.

Keywords: FIREX-AQ; active fire and smoke detection; big data applications; clustering; computer vision; image segmentation; self-supervised deep learning; multi-modal data fusion

1. Introduction

An important and common application of machine learning (ML) is to identify and leverage latent patterns in data or imagery. A typical approach is to use *supervised* learning, which requires a set of truth labels that the ML method attempts to generalize to the problem of mapping from an input dataset X to the output Y through a set of features, M. The challenge

with supervised learning, and even the recently popularized semi-supervised learning, is acquiring a sufficiently large and **unambiguous** set of labels, which often requires many hours of manual labor on the part of domain experts. Alternatively, *self-supervised* learning takes a similar input dataset X and finds relationships among the features M resulting in context-free groupings in the output Y. Because no labels are provided for the input, there are no labels provided in the output. To utilize the results, the labels or missing context must be assigned after the fact by experts, but this has proven to be a much less labor-intensive endeavor, all while keeping subject matter experts in the loop.

In previous work, we demonstrated that feeding 2-dimensional images of instrument radiances, or Level 1 (L1) data, into Deep Belief Networks (DBNs) coupled with an unsupervised clustering method results in images automatically segmented into relevant geophysical objects [1]. We further demonstrated that the same results can be achieved using a simplified architecture across select areas of the globe and for various kinds of land surface and atmospheric segmentation tasks [2].

In our recent work [3], we have generalized our ML framework into an open-source software system called Segmentation, Instance Tracking, and data Fusion Using multi-SEnsor imagery (SIT-FUSE). This framework allows for various types of encoders, including regular and convolutional DBNs, Transformers, and Convolutional Neural Networks (CNNs), and we have moved from traditional unsupervised clustering to a deep learning-based clustering approach.

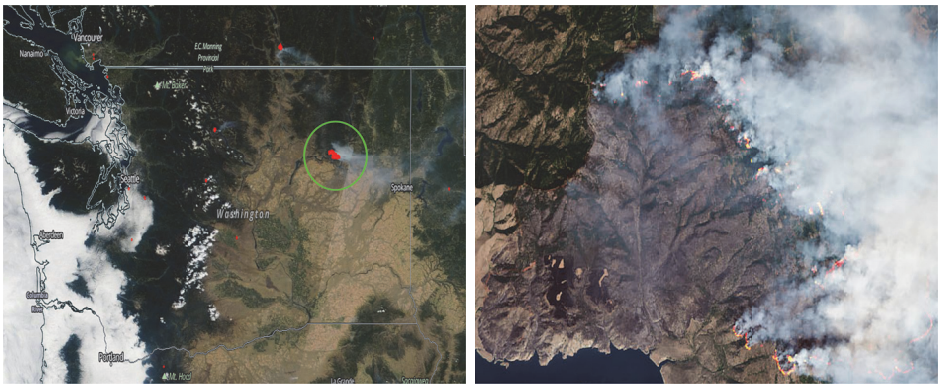
This approach, as a whole, has several unique benefits. First, it is not restricted to a particular remote sensing instrument with specific spatial or spectral resolution. Second, it has the potential to identify and "track" geophysical objects across datasets acquired from multiple instruments. Third, it allows for the joining of data from different instruments, "fusing" the information within the self-supervised encoder. Finally, it can be applied to many different scenes and problem sets, most notably in no- and low-label environments, not just ones for which labeled training sets exist, which is required for strictly supervised ML techniques.

Here, we apply our self-supervised ML approach to the problem of automatically detecting and tracking active wildfires and smoke plumes, through sequences of open-access L1 (imagery) data acquired by multiple remote sensing instruments during the joint National Aeronautics and Space Administration/National Oceanic and Atmospheric Administration (NASA/NOAA) Fire Influence on Regional to Global Environments and Air Quality (FIREX-AQ) field campaign that took place in the western United States in the summer of 2019 [4]. The high-altitude NASA ER-2 carried seven remote sensing instruments that provided high-spatial-resolution observations of active fires and smoke plumes in conjunction with NASA DC-8 aircraft and multiple satellite overpasses over the same fire events. The FIREX-AQ datasets of collocated satellites and multiple airborne imagery at different spatial resolutions are excellent as a testbed for the SIT-FUSE-based method of active fire/smoke identification and tracking, for which we have released the intermediate and final outputs for public access.

Wildfires and the smoke plumes induced by wildfires substantially contribute to the carbon cycle and can have a long-lasting impact on air quality and Earth's climate system. In addition, human-driven climate change is associated with more frequent and severe wildfires [5]. Despite the importance and immediacy of the problem, most research and decision-support tools to study wildfires and plumes use observations from a single instrument whose spatial coverage and (spatial, spectral, and temporal) resolutions vary from very fine to very coarse scales, neither of which, on their own, is fully capable of providing the much-needed information for a comprehensive understanding of wildfires and wildfire smoke [6]. As such, the current study aims to combine datasets with different spatial resolutions from multiple instruments to create a patchwork of datasets that fill

the temporal gaps present in current single-instrument active fire detection datasets. Here, the first step is testing a general framework for segmenting the datasets from multiple instruments and identifying wildfires and smoke plumes. Figure 1 shows a map of the active fire (red area within the green circle), taken from NASA's WorldView Snapshots web tool and a close-up reference image of the Williams Flats fire, one of the fires we focus on within this study, taken from the Landsat-8 Operational Land Imager (OLI).

The detection and tracking of objects, like wildfires and smoke plumes, within a single-instrument dataset has long required developing instrument-specific retrieval algorithms. Such development is labor-intensive and requires domain-specific parameters and instrument-specific calibration metrics, alongside the manual effort to track retrieved objects across multiple scenes [7]. The recent development of retrieval algorithms is actively underway in the field of supervised deep learning (DL), and various methods (e.g., Convolutional Neural Networks (CNNs)) have been applied. Some of these DL methodologies work well, in terms of precision and accuracy [7], but are still limited by the requirement that the spatial resolutions between training datasets and output products be the same. These methods also require pre-existing label sets, unlike recent supervised approaches like Fully Convolutional Networks (FCNs), Mask R-CNNs, and Transformers [8–10], which require large label sets to archive accurate results.



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

Figure 1. Reference map and imagery. (**a**) Map of fire location from publicly available NASA WorldView Snapshots/MODIS [11,12]. (**b**) Publicly available reference Williams Flats fire image from Landsat-8/OLI [13].

In our previous work, we demonstrated that an encoder trained in a self-supervised manner, namely a Deep Belief Network (DBN), trained with L1 (instrument radiance) images, can segment images based on geophysical objects within the scene, in conjunction with unsupervised clustering [1]. The unique benefit of this method is that its application is not limited to a single spatial or spectral resolution, and the method has the potential to detect and track objects from images with different resolutions from multiple instruments. With this method, instead of requiring a per-instrument finely hand-labeled label set, we can apply a coarser manual context assignment after segmentation on a smaller set of training scenes, allowing for this technique to be easily applied in cases of no labels or limited labels. We have also quantitatively validated that the same could be achieved using a simpler architecture for a set of atmospheric and land surface classification tasks using varying spectral, spatial, temporal, and multi-angle remote sensing data as input [2]. Since this work, we have transitioned from unsupervised clustering to self-supervised deep clustering, which we will discuss further in the Section 2.3. This completely self-supervised approach can leverage training data from many different scenes, not just ones that are accounted for by previous label sets for training, as is the case with strictly supervised

techniques. Ongoing research applies this self-supervised machine learning methodology to track detected smoke plumes across spatiotemporal domains. However, this study focuses on identifying wildfire and smoke plumes within a single-instrument dataset and using a fusion of datasets from multiple instruments.

This approach not only allows us to leverage single- and multi-instrument datasets to create a denser static patchwork of active fire and smoke detections with increased spatial, spectral, and temporal resolution (as depicted in Figures 2–4), but it also gives us a uniform embedding-based representation of the data via the encoder outputs and final output of clusters. The final cluster output can be used in conjunction with spatial distributions of the output labels to facilitate active fire and smoke plume instance tracking across multi-sensor scenes over varying spatiotemporal domains. Figures 2–4 demonstrate the various tiers and scales of representative capabilities over the Williams Flats fire on 6 August 2019, when incorporating observations GOES at the coarse spatial but fine temporal resolution end of the scale, and the airborne instruments mentioned in Table 1 over the Williams Flats and Sheridan fires at the fine spatial but coarse temporal end of the scale, along with the polar orbiters in-between these two extremes.

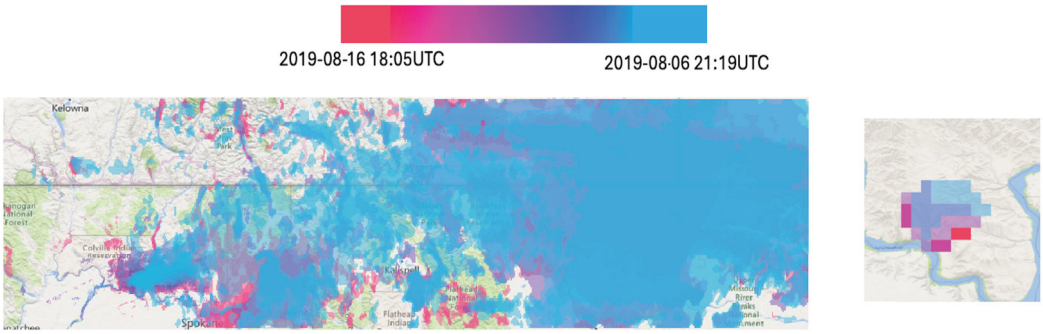


Figure 2. A display of the smoke plume (**left**) and the active fire front (**right**) progression of the Williams Flats fire on 6 August 2019, as captured by GOES-17 and segmented via SIT-FUSE.

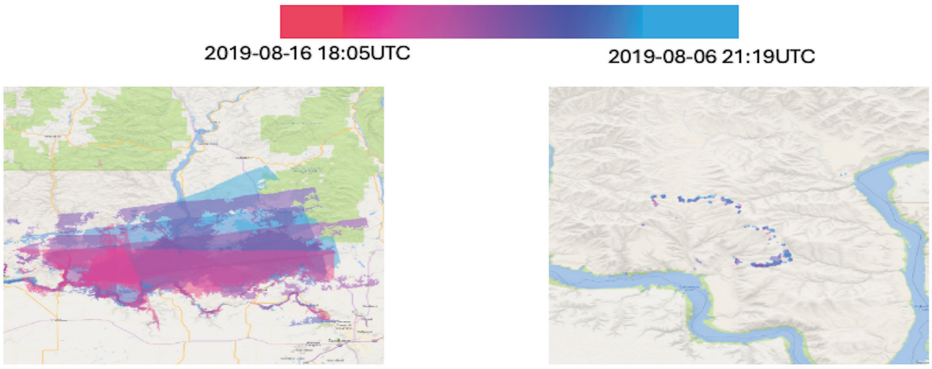


Figure 3. A display of the smoke plume (**left**) and the active fire front (**right**) progression of the Williams Flats fire on 6 August 2019, as captured by eMAS and segmented via SIT-FUSE. When used in conjunction with the data in Figure 2 we can look across large spatial domains at fine temporal scales as well as fine-scale detail at higher spectral resolutions.

Work on the general problem of self-supervised image segmentation appears to have had success in separating the foreground from the background [14,15], or have made significant subsets of spectral resolution (using a single band of input) from one type of instrumentation, which is effective for their applications, but does not provide the spectral specificity or per-observation or temporal resolution we are trying to attain here [16]. Other works have focused on urban planning and mapping, outlining buildings and roadways [17], which is not the goal here. A similar study that used a similar machine learning

approach to us—using autoencoders for representation learning and clustering for unsupervised segmentation—attained an accuracy of 83% on Landsat imagery alone [18]. This uses a similar kind of model to our studies but uses a single instrument. With large variations in spatial and spectral resolutions, our technique attains higher accuracy (and balanced accuracy, in some cases) across many different instrument sets, including fused data. Even with more recent breakthroughs in semi-supervised semantic segmentation, like the Segment Anything Model (SAM), a problem-dependent amount of labels is required, and SAM is largely unproven in complex domains like remote sensing [19]. The identification of the necessary size of label sets, generation of per-pixel label sets, and testing of the feasibility of new techniques in more complex domains are all problem-specific and time-consuming tasks that can be skipped, given our solution—as seen in the successful but extremely limited cases discussed in [20–23]. Lastly, there are new physics-based retrieval techniques, which seem promising, but need continued rigorous analysis to generalize across different regions and instrument types [24]. In the future, it may be useful to combine the physical parameterizations and ML-based retrievals via ML loss functions that are "physics-aware". The lack of need for large new label sets mitigates the costly, labor-intensive work of manually segmenting each pixel within a dataset used for ground truth, a process which is itself error-prone, and other previously mentioned supervised learning-related precursors model training. Also, leveraging pre-existing operational products to use as labels for supervised learning tasks will inherently cause them to either lack training set diversity or suffer from the issues mentioned above. On the other hand, our approach is well suited to handle large amounts of data, because our unsupervised and self-supervised models can perform label-free image segmentation. The fact that the human-in-the-loop steps of context application and validation occur after the images have been segmented allows for human oversight while mitigating the need for the extremely labor-intensive act of pixel-by-pixel manual segmentation for tens of thousands of images. In the subsequent sections, we will describe the experimental design for evaluating the performance and efficacy of using SIT-FUSE in support of Fire Influence on Regional to Global Environments Experiment—Air Quality 2019 (FIREX-AQ 2019; https://csl.noaa.gov/projects/firex-aq/, accessed on 20 December 2024), the results, the conclusion of the experiments, provide further discussion points, and discuss current and future work on this approach, the associated framework, and the correlated tooling.

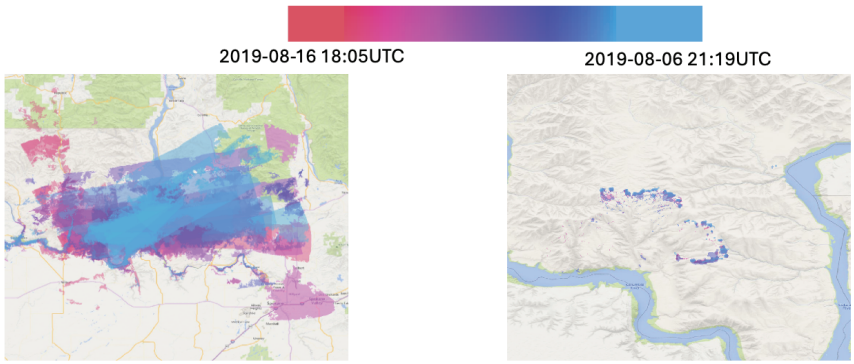


Figure 4. A display of the smoke plume (**left**) and the active fire front (**right**) progression of the Williams Flats fire on 6 August 2019, as captured by eMAS, MASTER, AirMSPI, and AVIRIS-C and segmented via SIT-FUSE. This combination of instrumentation maximizes the temporal resolution at the associated high spatial and spectral (/polarization) resolutions of the airborne instrumentation. When compared to the data in Figure 3, this increases the yields even more for monitoring and science capabilities when used in conjunction with geospatial and polar orbiting instrumentation.

Table 1. Airborne instruments and their products.

Platform	Instruments	Science Products	Spatial Resolution
NASA ER-2	Airborne Multiangle SpectroPolarimetric Imager (AirMSPI) [25]	Spectro-polarimetric intensities (10 m spatial resolution, 8 wavelengths in 355–935 nm spectral range, 3 polarimetric bands)	10 m
NASA ER-2	Enhanced MODIS Airborne Simulator (eMAS) [26]	Spectral intensities in 38 bands in 445–967 nm and 1.616–14.062 µm spectral ranges	50 m
NASA DC-8	MODIS/ASTER Airborne Simulator (MASTER) [27]	Spectral intensities in 50 bands in 0.44–12.6 µm spectral range	10–30 m
NASA DC-8	Airborne Visible/Infrared Imaging Spectrometer—Classic (AVIRIS-C) [28]	Spectral intensities in 224 bands in 400–2500 nm spectral range	10–30 m

2. Materials and Methods

2.1. Input Data

To demonstrate the efficacy of the approach for multi-platform smoke and active fire front segmentation for field campaign support, we used remote sensing datasets from airborne and satellite instruments collected during the FIREX-AQ 2019 campaign. Specifically, we investigated the wildfires and smoke plumes originating from the Williams Flats fire, Horsefly fire, and Mosquito fire over the three days from 6 August through 8 August 2020. Where available, we also tested on scenes over the Sheridan Fire, from 15 August through 23 August. The combination of different fires and associated landscapes for the central fires of the campaign (Williams Flats and Sheridan) as well as two additional fires (Horsefly and Mosquito) give us geospatial, land cover, and temporal variance to evaluate over while demonstrating efficacy in the field campaign setting. The FIREX-AQ campaign involved two aircraft: the DC-8 aircraft with a primary payload of in situ instruments supplemented by several remote sensors, and NASA's high-altitude ER-2 aircraft with a package of seven remote sensing instruments. The ER-2 aircraft coordinated on a few occasions with the DC-8 aircraft and had multiple collocations with satellite sensors flying directly along the satellite track. Therefore, several observations of the same active fires and smoke plumes were made by multiple instruments at various spatial resolutions. Tables 1 and 2 summarize the datasets from airborne and satellite instruments used in this study.

Table 2. Satellite instruments and their products.

Platform	Instruments	Science Products	Spatial Resolution
Terra	Multi-angle Imaging SpectroRadiometer (MISR) [29]	Spectral intensities at 446 nm, 558 nm, 672 nm, and 867 nm	1.1 km and 275 m, all resampled to 1.1 km
Terra and Aqua	MODerate resolution Imaging SpectroRadiometer (MODIS) [11]	Spectral intensities in 38 bands in 445 nm–967 nm and 1.616 µm—14.062 µm spectral range	1 km when used alone and resampled to 1.1 km when used with MISR
Suomi NPP and NOAA-20	Visible Infrared Imaging Suite (VIIRS) [30]	Spectral intensities in 5 bands in 0.44–12.6 µm spectral range	375 m
GOES-17	Advanced Baseline Imager (ABI) [31]	Spectral intensities in 50 bands in 0.47–13.3 µm spectral range	1 km and 2 km, all resampled to 2 km
PlanetScope	Dove Imagers [32]	Spectral intensities in 4 bands in 455–860 nm spectral range	1–3 m native resolution, all resampled to 10 m per the NASA Technical Report https://ntrs.nasa.gov/citations/20240001694, accessed on 10 January 2025

2.2. Data Preprocessing

For single-instrument cases, data were re-projected to the WGS84 Latitude/Longitude projection. For multi-sensor cases, data are collocated, re-projected to the WGS84 Latitude/Longitude projection, resampled to the lowest common spatial resolution,

and stacked channel-wise. The actual fusion occurs as a part of the representation learning performed inside the encoder.

Samples were generated by taking pixels and their direct neighbors (center pixel + 8 neighbors), in all channels, and creating a flat vector. These samples were then standardized by removing the per-channel mean and scaling to per-channel unit variance. The per-channel mean and standard deviations are computed over the full sample set from the scenes used for training.

For training, the scenes used are subsets. To ensure the subset of samples from the training scenes contained a representative variety of terrain and phenomena, k-means clustering was applied to the data with a set of 50 classes. This stratification technique was chosen as it has proven to be an effective, albeit naive, way to ensure variation within training samples [33–35]. Concerning the number of classes chosen (50), we ran a simple elbow criteria analysis, identifying the number of classes that maps to the beginning of diminishing returns relative to the inertia, or the sum of squared distances between each point and the centroid of the cluster it has been assigned to. Figure 5 depicts the cluster count vs. inertia plots for each instrument. Three million samples are randomly sampled based on the stratification generated by the full number of samples from the training scenes labeled with the 50 k-means classes. All pixels that are set to fill values, or out of specified valid ranges are discarded before any preprocessing. Regarding the spectral bands used, all spectral bands were used except bands that were extremely noisy or known to be non-functional for the FIREX-AQ campaign time frame. The same pixels used to train the encoders are also used to train the deep clustering heads, and the full training scenes are used to assign context.

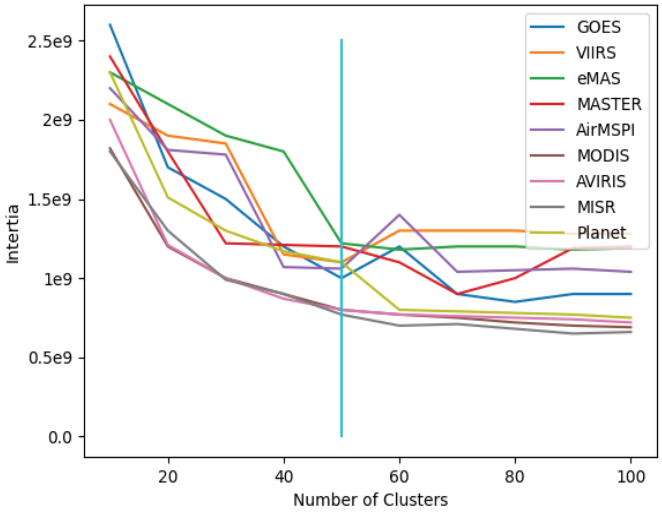


Figure 5. A set of the cluster count vs. inertia plots for each instrument that informed our selection of the number of clusters to use for the k-means-based naive data stratification approach we employ. The horizontal line is at K = 50, the selected number of clusters.

SIT-FUSE can take larger tiles for Convolutional DBNs, CNN-based, and Transformer-based architectures (as mentioned below). However, the pixel neighborhood has proven to provide enough spatial context with regular DBNs, as used in this work. Figure 6 is a flow chart that depicts the overall flow of SIT-FUSE.

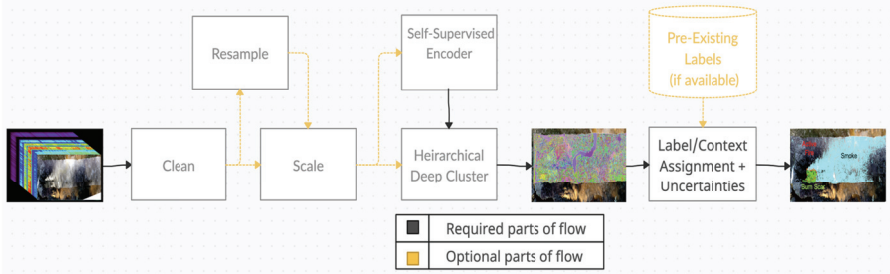


Figure 6. A flow diagram for the processing of one input type (single-instrument or fusion set) through SIT-FUSE.

2.3. Methods

2.3.1. Self-Supervised Representation Learning

SIT-FUSE is developed to be a generic framework allowing various kinds of encoders and foundation models that leverage self-supervised representation learning, including Deep Belief Networks (DBNs) trained using contrastive divergence, Convolutional Neural Networks (CNNs) with residual connections trained via Bootstrap Your Own Latent (BYOL), and Transformers trained using Image-Joint Embedding Predictive Architecture (I-JEPA) or Masked AutoEncoders (MAEs), as well as pre-trained Transformers for Earth Science, like Clay (https://madewithclay.org/, accessed on 10 November 2024) [36–39]. For all of these experiments, we used DBNs with 2–3 layers. DBNs were selected here because previous work and experiments performed in this work demonstrated that they produce reasonable results and the parameter space is much smaller than the other models mentioned above. We have performed extensive validation on the use of DBNs from both the perspective of structural understanding and downstream task performance, as well as resource consumption assessment, for the large set of single-instrument and fusion datasets [2]. In short, 2–3 layer DBNs provide a relatively compact model (2 million parameters vs. 100 million–10 billion parameters) with representational capabilities that meet our needs. While given the per-layer training paradigm and the generative nature that has caused many to turn to other architectures/approaches, recent work has demonstrated out-sized representational capabilities relative to other much larger generative models, like Generative Adversarial Networks (GANs) [38]. Given our successes and studies like [38], we continue to provide DBNs as an encoder option within SIT-FUSE, while also acknowledging the difficulties others have had with them in the past. We are currently evaluating encoder complexity in relation to segmentation performance and geographic coverage, to optimally operationalize this approach for operational global production, which will be discussed further in later manuscripts.

As in our previous work [1–3], the DBN architectures used leverage feature expansion, outputting embeddings of a pixel neighborhood in a larger feature space than the one with which the samples were input. Similar to the idea behind input kernelization, architecture-based feature expansion allows the models to learn nonlinear latent patterns that would be compact and complex in lower dimensions in simpler but higher-dimension forms. This method, although not the most common use case for DBNs or encoders in general, has been demonstrated effectively in other studies as well as in our previous work [1,2,40–42]. In previous works, we held architecture parameters static for all tests to demonstrate efficacy. Here, we varied the number of layers and hidden/output parameters for each encoder used, based on input data spectral resolution.

2.3.2. Deep Clustering

To extract segmentation maps from the per-pixel embeddings, we use deep clustering, specifically Information Invariant Clustering.

Previous experiments used BIRCH and other forms of traditional agglomerative clustering. We have transitioned to DL-based clustering because the training time, inference time, memory requirements, and model re-usability are all much improved when using neural network layers via PyTorch2.0 when compared to using traditional clustering via sci-kit-learn. These layers are trained using the invariant information clustering (IIC) loss function. IIC aims to assign labels that maximize mutual information between an input sample x and a perturbed version of x, x' [43]. For our use, perturbations are additions of Gaussian noise to the outputs of RBM-based architectures.

To mimic the hierarchical nature of agglomerative clustering output, we have set up hierarchical deep clustering layers. Here, the output heads are set up in a tree structure where each sub-tree is only trained on label samples classified as belonging to their parent label sets. Each layer receives only the output of the encoder, but as the tree is built from the top down, each neural network in a child node position on the tree only sees samples assigned labels associated with its parent and ancestor node(s). In this way, we can create scene segmentation at varying levels of specificity and explore the connections between each level. To our knowledge, this is the first study/software system that leverages IIC layers, or even deep clustering layers, in this fashion. For each of the pipelines, the root IIC head allows for 800 possible classes and each child IIC head allows for 100 subclasses. Currently, the level of hierarchy is specified by the user and there is no automated node splitting, but this could be implemented in the future. For our purposes, only two levels of hierarchy are used. We believe that this not only allows us to attain the specificity required for our segmentation tasks but that the hierarchical labeling can be leveraged as a co-pilot for data exploration and discovery. Figure 7 shows a two-level tree/flow diagram for performing deep clustering.

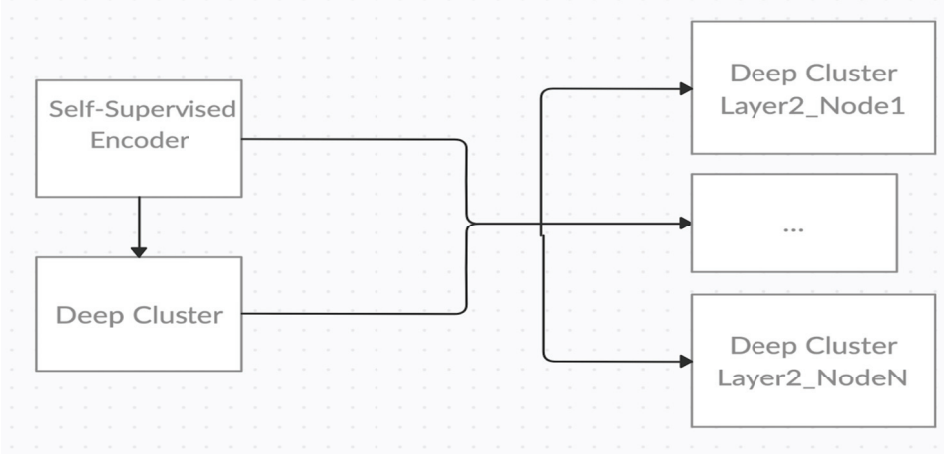


Figure 7. A 2-layer flow diagram example of the setup for hierarchical deep clustering. Each box labeled 'Cluster' is a set of fully connected layers, connected to the RBM-based model and trained via the IIC loss function. Each child node is only trained and makes predictions on samples given the label from its parent nodes that are associated with the path of edges that link the root to the current node. This setup allows us to use deep clustering to create interlaced levels of specificity for data exploration and characterization.

2.3.3. Hand Labeling

For both context assignment and validation, subject matter experts labeled areas of high-certainty smoke, active fire, and associated backgrounds. All areas that labelers were uncertain about, remained without labels. This labeling process was performed by generating polygons over the remote sensing imagery. An example is provided below in Figure 8. Because scenes can have overlapping classes (i.e., active fire and smoke are contained in the same pixel), but also have areas distinct to a single class, a separate

background class label set was generated for active fire and smoke. Figure 8 illustrates the labeling process over a single AVIRIS-C scene.

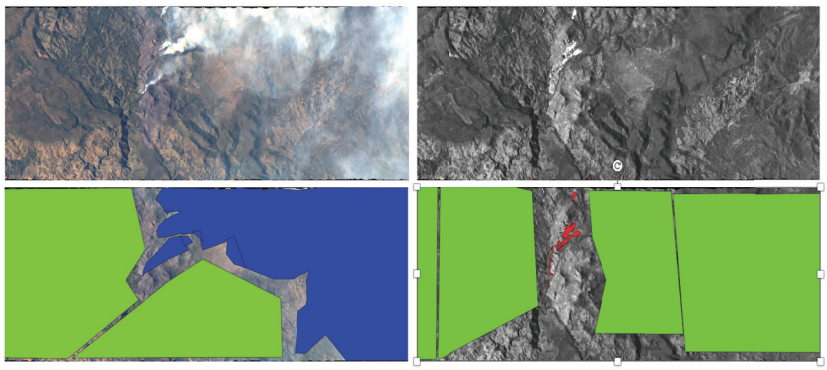


Figure 8. Visualization of the labeling process. The top row depicts a reference RGB generated from an AVIRIS-C scene (**top left**) and a single thermal band used for identifying active fire front locations (**top right**). The corresponding image in the bottom row has the smoke and smoke background labels over the RGB image and the active fire and fire background labels over the thermal image.

Unlike labeling for supervised learning, this approach does not require all training samples to be labeled, which is relevant for problems with high uncertainty in boundary cases, like the segmentation of active fire fronts and smoke plumes. The labeling of only the high-certainty class areas allows us to capture and compare against segmentation structure, and provide ample samples for context assignment. Although this labeling minimizes the pre-analysis labor required from subject matter experts, it still keeps experts in the loop (a crucial piece for science-related ML tasks). Relative to the number of scenes labeled here, supervised and semi-supervised tasks require 100× more labeled samples or more. Also, because they are learning the mapping between the labels and the input datasets directly, they require much more complete label sets (i.e., uncertain areas must be labeled background or foreground, potentially leading to systematic over-segmentation or under-segmentation).

2.3.4. Context Assignment

To assign context to the context-free segmentation maps generated via SIT-FUSE, we used zonal histogramming. For this, we overlayed the labels for a specific scene on the segmentation map generated for the same scene and generated counts of overlapping occurrences of the segmentation map classes and the high-certainty smoke, active fire, and background labels. These were gathered over multiple scenes, and then each class in the segmentation maps was assigned to the labels it best matched. This assignment was performed separately for each class of interest and its associated background class to ensure multi-class representation, like pixels containing both smoke and active fire. Figure 9 provides an example of labels overlayed on a segmentation map.

2.3.5. Contouring and Filtering

As the pixels within the detection are components of larger, often connected, geophysical objects, SIT-FUSE allows for building objects from separate detection pixels. To achieve this, we leverage the contouring capabilities within the 2.0 openCV Python/C++ software package [44,45]. Once generated, these contours are filled and we are left with multiple separate objects, instead of many more distinct pixels. We can also filter out small and large objects, where appropriate, and while not used here, leveraging dilations and erosions on the contoured objects is also possible. To mimic the automated process of an opera-

tionalized version and ensure fair comparisons and validation, all processing is performed uniformly across all scenes in an input set.

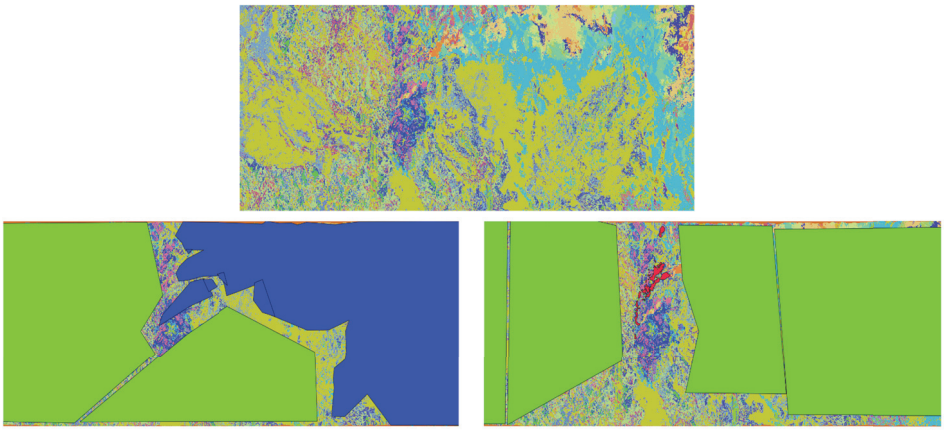


Figure 9. Visualization of the overlay required for the zonal histogramming process. The top row contains the context-free segmentation map generated via SIT-FUSE for the scene in Figure 8. The bottom row contains the same map with the smoke/background (**left**) and active fire/background (**right**) labels overlayed. The zonal histogramming process provides counts of classes in the segmentation map relative to label positions and each class is assigned to the label it best agrees with over multiple scenes.

2.3.6. Validation

There is no direct ground truth here—meaning comparisons are performed against pre-existing detections that are both used operationally and are considered the operational state of the art, as well as hand labels generated over only the high-certainty areas of smoke plumes and active fire fronts—and we know there are other areas of these objects that are not included in the labels. We include both comparisons as part of the validation because neither dataset is actual ground truth, and both comparisons provide valuable perspectives on performance. For each type, we split out the experiments and evaluation metrics to provide a clear picture of performance relative to the two different kinds of ground-truth-proxies used for performance evaluation in this study. Given the fact that there are areas known to be omitted from both datasets we are comparing against, and areas of false positives in the pre-existing detections, recall, precision, and their collective summary via F1-scoring are too harsh of evaluators here. We evaluated some previously published versions of precision and recall that apply fuzzy logic, but ultimately landed on the structural similarity index (SSIM) to evaluate performance across the various dataset pairs [46,47]. This is a fairly common problem within the remote sensing domain and one we aim to help solve with the collective incorporation of self-supervised learning, subject matter expert domain knowledge, and large amounts of data [48,49].

2.4. Materials and Tools

The software was developed with Python 3.9.13. SIT-FUSE has open-source functionality at its core [50]. To achieve the required goals of the software and leverage pre-existing and well-validated open-source software, geospatial, big data, and ML toolkits are the backbone of SIT-FUSE. For optimized handling and computation on large datasets across CPUs and GPUs, numpy, scipy, dask, xarray, Zarr, numba, and cupy are used [51–57]. For CPU- and GPU-based ML model training, deployment, evaluation, and auto-differentiation, sci-kit-learn, PyTorch, and torchvision are used [58,59]. Because RBMs are not included within the PyTorch library, Learnergy, an open-source library that contains various PyTorch-backed RBM-based architectures is used as well [60]. On the geospatial side of the problems being solved,

pyresample, GDAL, OSR, healpy, polar2grid, and GeoPandas are leveraged [61–65]. Lastly, for non-machine learning computer vision techniques, OpenCV is used [66]. The combination of these commonly used and well-tested software systems allows us to employ state-of-the-art approaches and architectures with minimal development and maintenance efforts, most of which are only minimally visible to the end user. SIT-FUSE is also publicly available and maintained on the public version of GitHub. For labeling, context assignment, and visualization/qualitative assessments, QGIS2.2, an open-source Geographic Information System (GIS) was used [67].

The hardware utilized was an NVIDIA GeForce Titan V100 GPU with 32 GB memory, as well as the NCCS Prism GPU Cluster (https://www.nccs.nasa.gov/systems/ADAPT/Prism, accessed on 7 October 2024).

3. Results

3.1. Fire Detection

Table 3 summarizes the performance of our SIT-FUSE-based segmentation approach, compared against the hand-labeled high-certainty active wildfire front areas, for active wildfire fronts across all scenes in the test set, for each instrument tested. Table 4 presents the same, but only for active fires and geographic areas not seen during training. In all cases, our approach performs well. When visually compared, our approach tends to over-segment relative to the high-certainty active fire pixels. However, the vast majority of the over-segmentation is over areas burning at lower heat, which are therefore less visibly associated with the active fire front. This effect can be seen for the Williams Flats fire in Figure 10 and for the Sheridan fire in Figure 11.

Table 3. Summary of active fire detection comparisons against the hand-labeled high-certainty active wildland fire fronts over the full set of test scenes that contain active fires. Total pixel count is the total number of pixels tested. SSIM was the metric used for comparisons in this study. For VIIRS and MODIS, the data from both platforms are used collectively.

Dataset	Total Pixel Count	SSIM
MASTER	12,438,023	0.73
eMAS	162,644,645	0.87
AVIRIS-C	177,233,951	0.89
GOES-17	1,016,785	0.82
VIIRS	4,759,386	0.88
MODIS	3,262,896	0.88

Table 4. Summary of active fire detection comparisons against the hand-labeled high-certainty active wildland fire fronts over only the scenes that contain the Sheridan fire and Horsefly fire, unseen during training. Total pixel count is the total number of pixels tested. SSIM was the metric used for comparisons in this study. For VIIRS and MODIS, the data from both platforms are used collectively.

Dataset	Total Pixel Count	SSIM
MASTER	1,708,746	0.89
eMAS	133,762,591	0.88
AVIRIS-C	21,073,834	0.93
GOES-17	2,541,975	0.84
VIIRS	1,189,846	0.88
MODIS	815,724	0.88

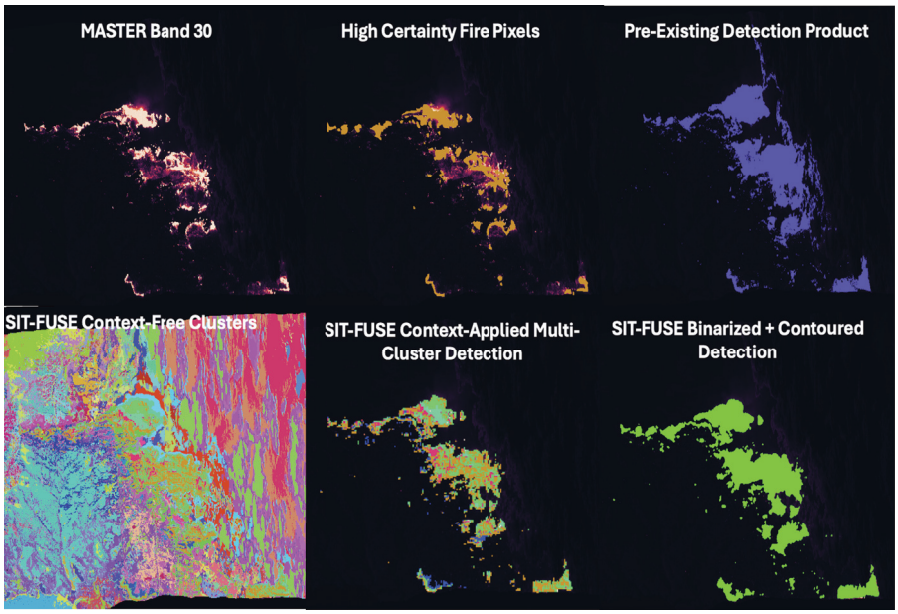


Figure 10. The top left panel is a single pseudo-colored thermal band from MASTER from a scene over the Williams Flats fire. Following this panel, on the top row are the hand-labeled high-certainty active fire pixels overlayed on the MASTER band, and the pre-existing band-ratio-based detection overlayed on the MASTER band, respectively. The bottom row consists of the context-free segmentation map generated from SIT-FUSE for this scene; the context-applied multi-cluster detection, which is a subset of the full segmentation map with only the labels/clusters that correlate to active fire fronts; and the final binarized and contoured active fire mask.

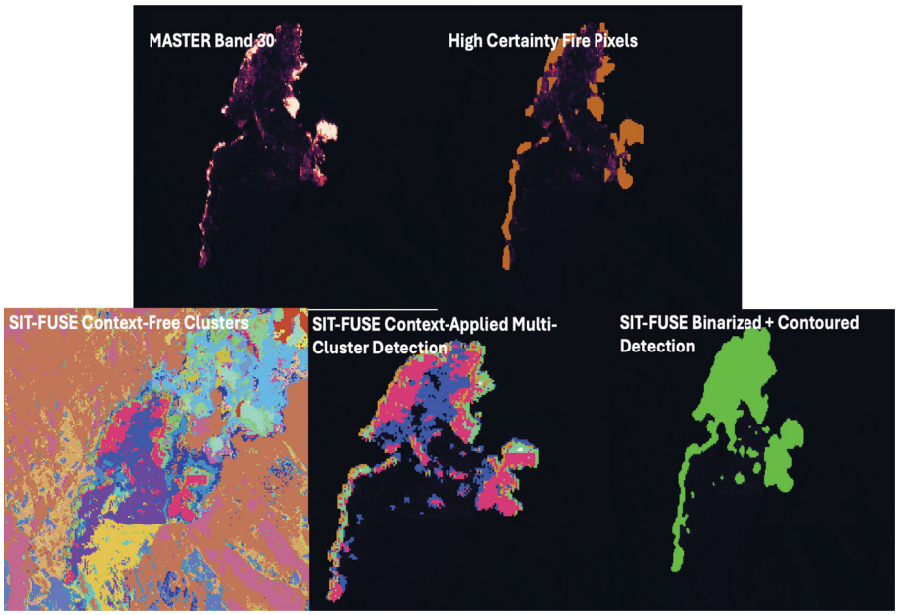


Figure 11. The top left panel is a single pseudo-colored thermal band from MASTER from a scene over the Sheridan fire. Following this panel, on the top row are the hand-labeled high-certainty active fire pixels overlayed on the MASTER band, and the pre-existing band-ratio-based detection overlayed on the MASTER band, respectively. The bottom row consists of the context-free segmentation map generated from SIT-FUSE for this scene; the context-applied multi-cluster detection, which is a subset of the full segmentation map with only the labels/clusters that correlate to active fire fronts; and the final binarized and contoured active fire mask.

Table 5 summarizes the same type of comparisons, this time against SIT-FUSE-based detections, and results from an experimental instrument-specific band-ratio-based detection methodology are compared against the high-certainty active fire pixel label set for MASTER.

For this comparison, we used all scenes in the test set with an associated band-ratio-based detection to compare against. The segmentation quality is quite comparable, with both techniques obtaining the same SSIM score; however, the difference maps in Figure 12 show a test-set-wide trend of hot smoke also being picked up as part of the active fire front in the band-ratio-based detection.

Table 5. Summary of active fire detection comparisons against the hand-labeled high-certainty active wildland fire fronts over only the scenes that contain the Sheridan fire and Horsefly fire, unseen during training. Total pixel count is the total number of pixels tested. SSIM was the metric used for comparisons in this study.

Dataset	Total Pixel Count	SSIM
MASTER SIT-FUSE	20,389,114	0.8
MASTER Pre-Existing	20,389,114	0.8
GOES SIT-FUSE	1,016,785	0.82
GOES Pre-Existing	1,016,785	0.72
VIIRS SIT-FUSE	4,759,386	0.71
VIIRS Pre-Existing	4,759,386	0.59

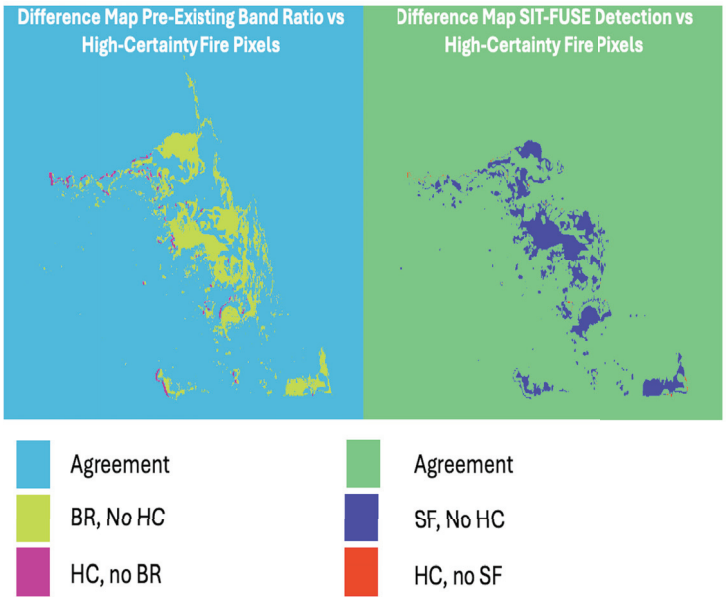


Figure 12. A pair of difference maps between different detection outputs and the hand-labeled high-certainty active fire pixels. The left panel is a difference map between the band-ratio-based detection (BR) and the high-certainty labels (HC). The right panel is a difference map between the SIT-FUSE-based detection (SF) and the high-certainty labels (HC).

3.2. Smoke Detection

Table 6 summarizes the performance of our SIT-FUSE-based segmentation approach, compared against the hand-labeled high-certainty smoke plume pixels, for plumes across all scenes in the test set, for each instrument tested. Table 7 presents the same specifically for smoke plumes in geographic areas not seen during training. In all cases, our approach performs well. When visually compared to the hand-labeled data, our approach tends to over-segment relative to the high-certainty smoke plume areas but performs well across the majority of the plume regions tested. The borders of the smoke plumes appear to provide uncertainty for both labelers and the SIT-FUSE-based automated detection. In some cases, these areas were not fully covered by our detections, whereas in other cases, the detections identified smoky areas in the scene that were not covered by the labels. Shadows also appear

to confuse small numbers of cases. Figure 13 provides an example detection produced for a MASTER scene capturing the Williams Flats fire, while Figure 14 shows the same example for an AirMSPI scene capturing the same active fire. Figure 15 shows difference maps over eMAS scenes of the same fire between the high-certainty hand labled data, SIT-FUSE smoke detections, and the eMAS AOD product, thresholded at 0.2 AOD. Our detection appears to capture more of the thick smoke plume, while not picking up areas surrouding the smoke plume that appear to be clear, but potentially have finer aerosols in the atmosphere.

Table 6. Summary of smoke detection comparisons against the hand-labeled high-certainty smoke plumes over the full set of test scenes that contain active fires. Total pixel count is the total number of pixels tested. SSIM was the metric used for comparisons in this study.

Dataset	Total Pixel Count	SSIM
MASTER	12,438,023	0.71
eMAS	162,644,645	0.63
AVIRIS-C	173,334,623	0.67
AirMSPI Sweep	132,238,386	0.63
PlanetScope Dove	1,353,594,326	0.68
GOES-17	1,016,785	0.62
VIIRS	4,759,386	0.66

Table 7. Summary of smoke detection comparisons against the hand-labeled high-certainty active fire areas over only the scenes that contain the Sheridan fire and Horsefly fire, unseen during training. Total pixel count is the total number of pixels tested. SSIM was the metric used for comparisons in this study. Planet data were not available over the Sheridan fire, so we cannot assess this.

Dataset	Total Pixel Count	SSIM
MASTER	1,708,746	0.75
eMAS	133,762,591	0.66
AVIRIS-C	21,073,834	0.77
AirMSPI Sweep	63,505,132	0.65
PlanetScope Dove	N/A	N/A
GOES-17	2,541,975	0.65
VIIRS	1,189,846	0.68

Table 8 summarizes the same set of comparisons, this time evaluating SIT-FUSE-based detections against results from operational dark-target-based detection methodologies. Both approaches are compared against the high-certainty active fire pixel label set for eMAS, GOES, and VIIRS. For these comparisons, we used the aerosol optical depth (AOD) value from the dark target product of each of the instruments and applied a minimum threshold value of 0.2.

Table 8. Summary of smoke detection comparisons against the hand-labeled high-certainty smoke plumes over only the scenes unseen during training. Total pixel count is the total number of pixels tested. SSIM was the metric used for comparisons in this study.

Dataset	Total Pixel Count	SSIM
eMAS SIT-FUSE	162,644,645	0.63
eMAS Pre-Existing	162,644,645	0.37
GOES SIT-FUSE	1,016,785	0.73
GOES Pre-Existing	1,016,785	0.56
VIIRS SIT-FUSE	4,759,386	0.66
VIIRS Pre-Existing	4,759,386	0.54

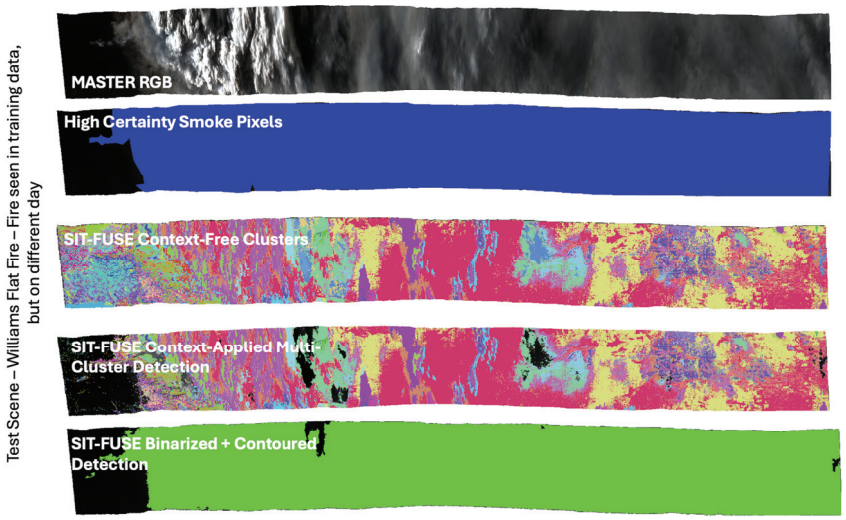


Figure 13. The top left panel is an RGB image extracted from a MASTER scene over the Williams Flats fire. Following this panel, on the top row are the hand-labeled high-certainty active fire pixels overlayed on the RGB image. The bottom row consists of the context-free segmentation map generated from SIT-FUSE for this scene; the context-applied multi-cluster detection, which is a subset of the full segmentation map with only the labels/clusters that correlate to smoke plumes; and the final binarized and contoured smoke mask.

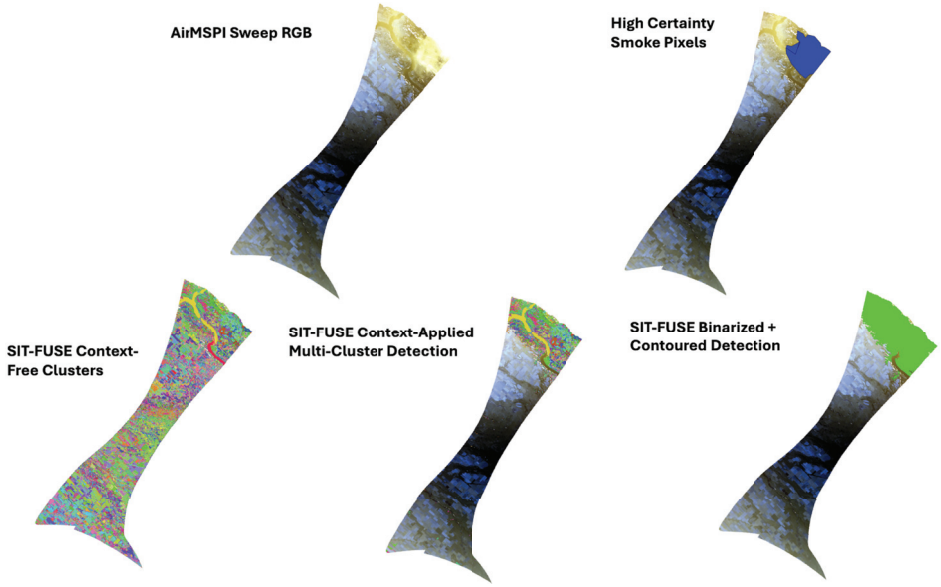


Figure 14. The top left panel is an RGB image extracted from an AirMSPI scene over the Williams Flats fire. Following this panel, on the top row are the hand-labeled high-certainty active fire pixels overlayed on the RGB image. The bottom row consists of the context-free segmentation map generated from SIT-FUSE for this scene; the context-applied multi-cluster detection, which is a subset of the full segmentation map with only the labels/clusters that correlate to smoke plumes; and the final binarized and contoured smoke mask.

3.3. Multi-Sensor Fusion

In previous work, we have demonstrated the utility of fusing multi-angle data from MISR with that of MODIS, both instruments onboard the same satellite platform, Terra. This included segmenting the Williams Flats fire and smoke plume in our evaluation. Figure 16 depicts segmentation, as well as active fire and smoke segmentation over the Williams Flats fire, using MISR and MODIS data as input. However, combining MISR and MODIS data is limited by the restricted number of overpasses for a single platform and the narrower area of swath intersection, which reduces the number of cases available for evaluation

compared to those in the single-instrument analyses described earlier. For polar-orbiting instruments, multiple platforms often provide an increased number of scenes for detection and evaluation. As we expand the spatiotemporal areas studied, we will add these fusion cases and quantify their benefits in a manner consistent with the previous evaluations.

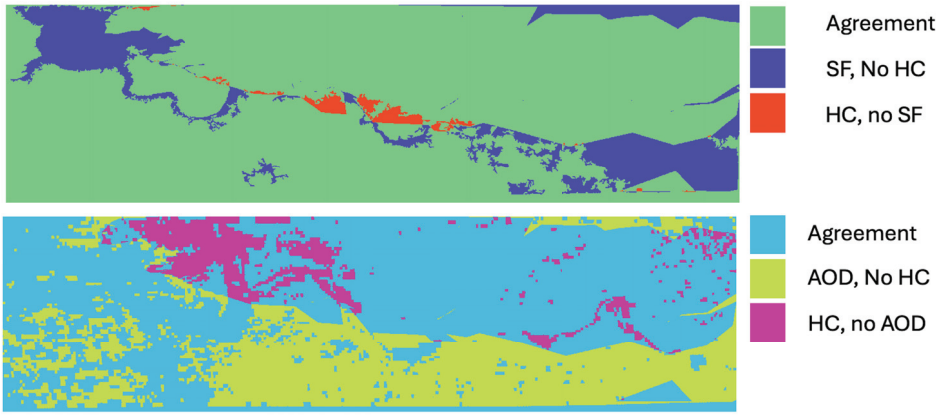


Figure 15. A pair of difference maps between different detection outputs and the hand-labeled high-certainty smoke pixels. The top panel is a difference map between the SIT-FUSE-based detection (SF) and the high-certainty labels (HC). The bottom panel is the difference between the operational eMAS AOD data, thresholded at 0.2 (AOD), and the high-certainty labels (HC).

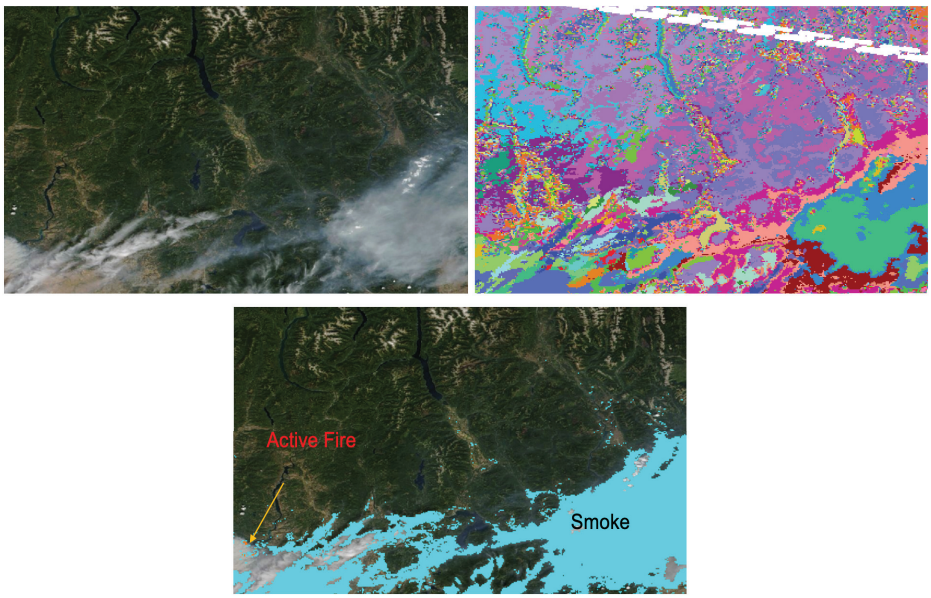


Figure 16. Example active wildfire front and smoke plume detection using Terra MODIS and 9-angle MISR scene over the Williams Flats fire on 8 August 2019, as input. The top left panel is a reference RGB generated from the MODIS data. The right panel is the context-free segmentation map generated from SIT-FUSE, and the bottom panel is the extracted and binarized smoke and active fire front detections from the SIT-FUSE outputs.

Given our previous success with the fusion of polar-orbiting instrument data, we explored fusing airborne instrumentation data, both within the same platform and across platforms. For the instruments on the same platform, we examined AirMSPI and eMAS, both onboard the ER-2 aircraft, as an airborne equivalent of MISR and MODIS. As a first step, we tested AirMSPI's multi-angle data taken in step-and-stare mode collectively as a simple fusion test case. Unfortunately, the coverage of the data is very small and the number of scenes is relatively low. While there is spatial overlap between data captured at

different viewing angles, data quality issues restrict their utility. There are areas marked as "bad/occluded data" that increased in frequency with more oblique angles. We tried to mitigate this by limiting the number of angles used, but it did not significantly improve, even after excluding data from four angles. Although we were able to generate smoke and active fire detection examples, the number of scenes and pixels was insufficient for both model training and separate evaluation. Figure 17 depicts an example scene from AirMSPI data, including the two most oblique angles and the nadir angle, as well as the minimal segmentation coverage we can obtain for SIT-FUSE when combining the multi-angle data, and the associated active fire and smoke masks.

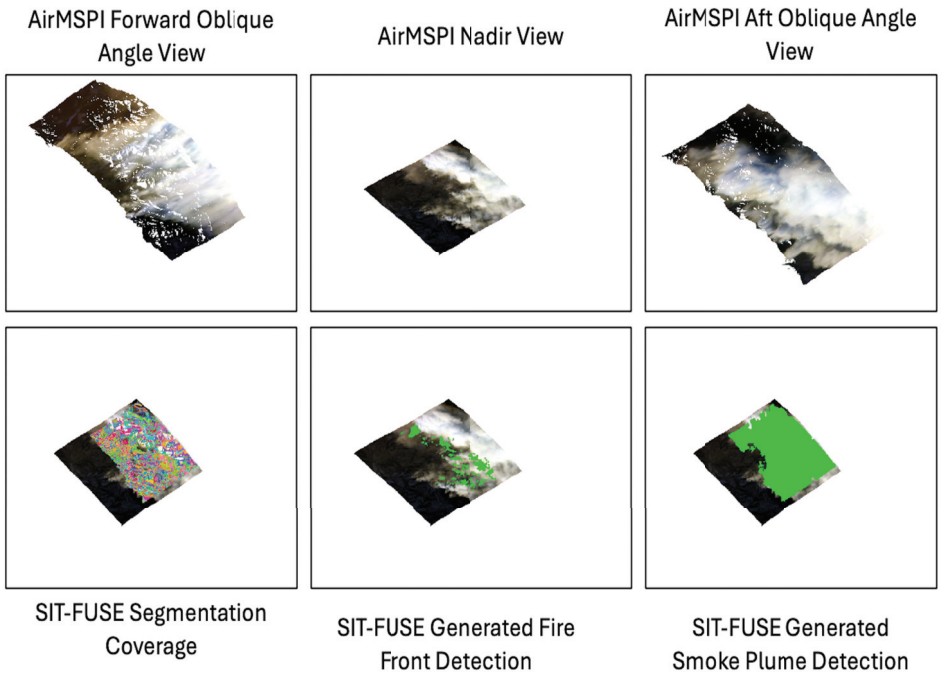


Figure 17. Example wildfire front and smoke plume detection using multi-angle AirMSPI step-and-stare data. The panels on the top row are RGBs from the two most oblique angles, split by the RGB generated from the nadir data from a scene over the Williams Flats fire on 6 August 2019. The bottom row depicts the full coverage for the SIT-FUSE segmentation map, overlayed on the nadir RGB for reference, the extracted active fire front detection, and the extracted smoke plume detection, respectively.

For the AirMSPI + eMAS case, we also attempted to combine the sweep mode data with eMAS data. However, the overlap between the datasets was minimal. Future campaigns may take into account the need to fuse data in this manner and attempt to collect a greater number of spatiotemporally collocated scenes with multiple instruments. Again, as we expand our case studies and datasets over larger areas and campaigns, we are confident that we can quantitatively assess the benefits of data fusion between AirMSPI and eMAS for active wildfire and smoke detection.

For the case of cross-platform fusion, we looked at fusing eMAS and MASTER. Again the number of scenes is limited, and only over the Williams Flats active fire area, but the initial examples demonstrate potential for this capability, and again future campaign planning and use of additional pre-existing campaign datasets will allow us to further validate these fusion cases. Figure 18 depicts an example of one of the fusion scenes.

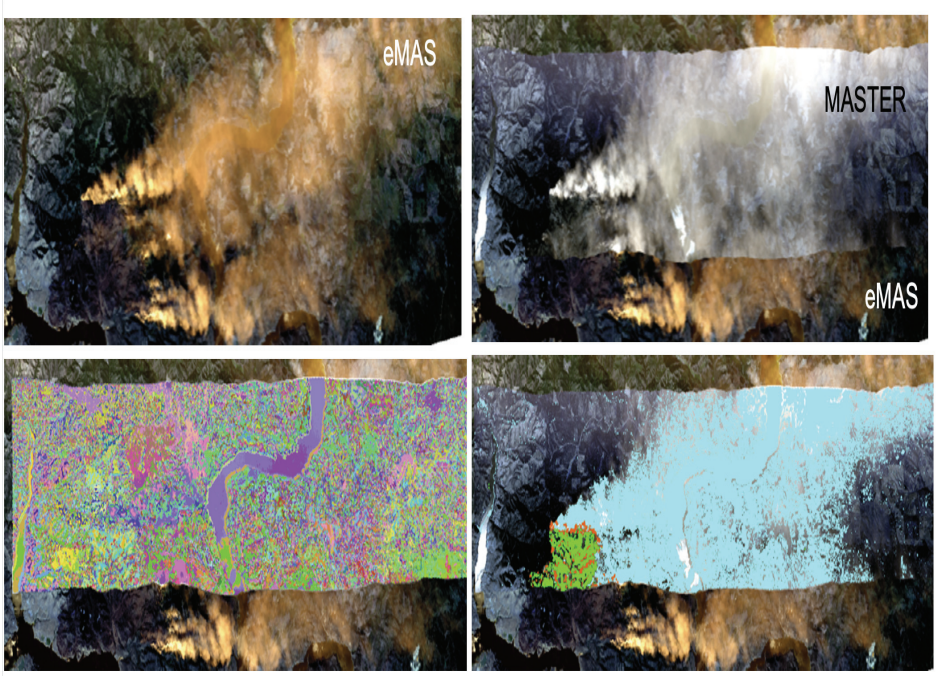


Figure 18. Example active wildfire front and smoke plume detection using MASTER and eMAS data as input. The top row depicts an RGB from an eMAS scene on 6 August 2019, and an RGB generated from the collocated MASTER scene overlayed on the eMAS RGB, both of which are used as input. The bottom row contains the context-free segmentation map overlayed on the area that contains both MASTER and eMAS data, and the extracted and binarized smoke and active fire front detections from the SIT-FUSE outputs.

3.4. Commercial Remote Sensing Products

As a part of this effort, we evaluated the utility of PlanetScope data for active wildfire and smoke detection and monitoring. We found that the older PlanetScope SuperDove imagers available during the FIREX-AQ 2019 campaign were not suited for identifying active fire fronts across the entire test set. However, they were effective at detecting smoke within the scenes evaluated. Figure 19 depicts smoke detection within a mosaic of SuperDove scenes over the Williams Flats fire on 8 August 2019, and Table 6 contains quantitative evaluation results. The improved temporal resolution of smoke detections at high spatial resolution will be helpful both for air quality research and automated smoke tracking. Also, as Planet's instrument offerings continue to diversify, with new SuperDove instruments having a higher spectral resolution, hyperspectral instrumentation being launched, and other companies continuing to launch various other Earth-orbiting instrumentation, it is critical that we look at the tradeoffs between paying for data from commercial entities and the scientific value added by such data. In this study, we analyzed 83 additional scenes over the Williams Flats area over four days. Per the NASA/Planet CSDSA agreement, we are not able to release the input SuperDove data, alongside all of the other data we are releasing. However, we will release the generated output and all other associated data.

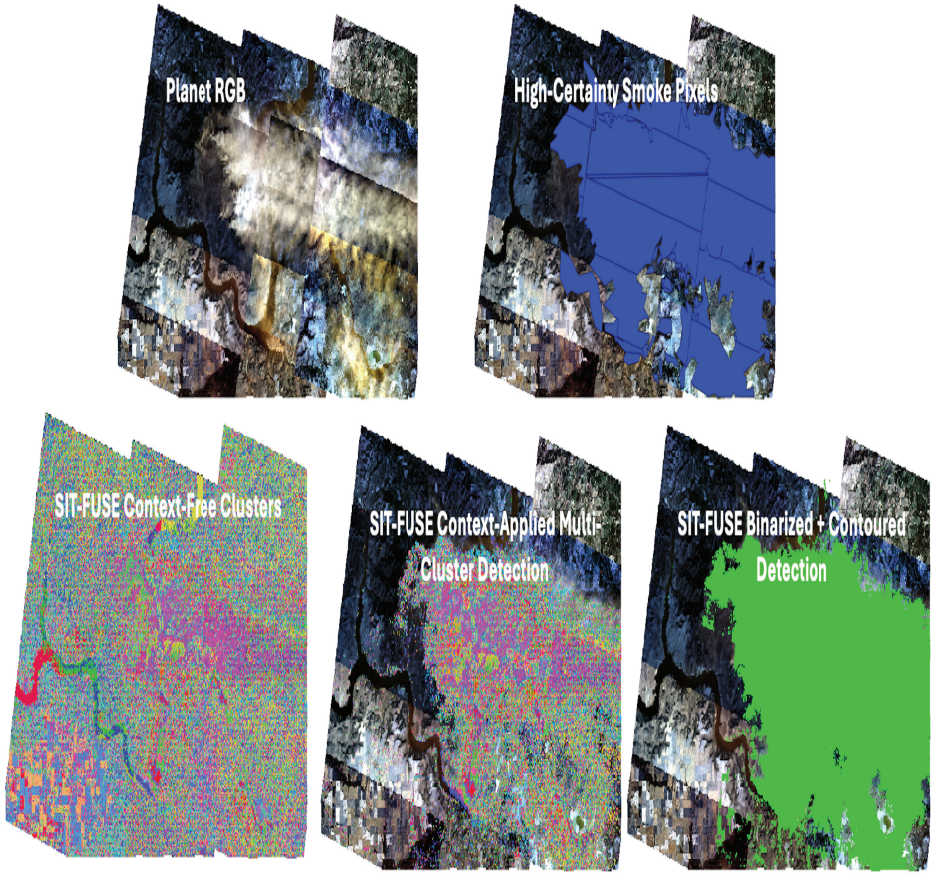


Figure 19. An example of using Planet data to perform smoke detection. The top row depicts a mosaic of RGBs generated from the input scenes and the high-certainty smoke pixels overlayed on top. The bottom row contains the context-free segmentation map generated from SIT-FUSE, the subset labels/clusters that are associated with smoke, and the final panel contains the contoured and binarized smoke plume detected over the entire mosaic. Image © 2019 Planet Labs PBC.

4. Conclusions

Overall, SIT-FUSE effectively identifies and segments wildland active fire fronts and smoke plumes, and performs better against high-certainty smoke and active fire label sets, when compared against other operational and experimental approaches, as seen in Tables 5 and 8. As expected, the higher the resolution of the instrument, the more detailed segmentation can occur. However, the different resolutions are valuable not only as standalone datasets but also for tracking these objects across multiple datasets. While further large-scale validation is needed, our tests demonstrate the capability to increase the temporal resolution of active fire front products. We achieve this by increasing the number of instruments available for active fire and smoke segmentation and developing brand-new active fire front and smoke plume products for many of the instruments tested. This indicates significant potential for both product generation and utilization of this technique for segmentation and instance tracking. This not only also allows for dynamic instance tracking across scenes with the same input set, but we believe by harnessing style transfer capabilities, we can also look into instance tracking across multi-sensor scenes from disparate input datasets. We hope to transition this technology into a piece of future operational campaign support or product generation systems.

5. Discussion and Current/Future Work

In terms of feature interpretability and selection, methods such as SHAP analysis and other explainability methods can be applied to better understand feature importance and refine the input to focus on spectral bands most effective for identifying smoke and/or active fire. Given the current performance and the success with datasets where there was no pre-existing operational active fire or smoke detection methodology, solutions like SIT-FUSE can be integrated into new or existing instrumentation data processing pipelines. By doing so, this approach could replace or augment instrument-specific retrieval algorithms, which may be extremely costly to develop. SIT-FUSE's segmentation capabilities offer additional benefits: the decrease in data volume processed for downstream active fire- or smoke-specific retrievals. By isolating the detected objects, only relevant pixels need to be processed through a downstream retrieval, thereby optimizing the pipeline.

We have built a framework within SIT-FUSE that is adaptable to various kinds of encoders and we aim to be able to leverage this to analyze representative capabilities of different model types, complexities, and training paradigms. With the continued influx of new architectures and large Earth Observation Foundation Models, it is important to understand these models provide quality representations (or poor ones) under different conditions, problem sets, and input datasets [68]. Analyses of downstream task performance are a crucial piece, but not the entire solution. More robust ways to evaluate representative capabilities are emerging around large language models (LLMs), and much of this can be ported to computer vision, and specifically deep learning for Earth observations [69]. Within the flexible framework of SIT-FUSE, we are working towards providing initial pathways towards tackling some of these open problems.

Lastly, we are working to leverage SIT-FUSE to make an impact within the area of analysis and scientific understanding—in this case, correlated to active wildfires and smoke plumes. There is a built-in co-discovery facilitation mechanism, by way of the hierarchical context-free segmentation products. By using the model-derived separations of various areas, novelty and "interesting" samples can more easily be grouped and investigated. This can be even further coupled with more detailed analyses of the embedding spaces relative to the context-free segmentations [3]. To enhance exploration even further, models trained for co-exploration of data using open-ended algorithms can be leveraged to more quickly sift through the volumes of data and highlight interesting, new, and anomalous samples [70,71].

Author Contributions: Conceptualization, N.L., K.Y., M.J.G. and E.L.; Data curation, N.L., A.E., H.L., M.J.G. and O.V.K.; Formal analysis, N.L. and A.E.; Funding acquisition, N.L., K.Y., H.L., M.J.G. and O.V.K.; Investigation, N.L., K.Y., M.J.G., O.V.K. and E.L.; Methodology, N.L., K.Y., H.L. and E.L.; Project administration, N.L.; Software, N.L., K.Y., H.L. and E.L.; Supervision, N.L., M.J.G., O.V.K. and E.L.; Validation, N.L., K.Y., H.L., M.J.G., O.V.K. and E.L.; Visualization, N.L., K.Y., H.L. and E.L.; Writing—original draft, N.L.; Writing—review and editing, N.L., K.Y., H.L., M.J.G., O.V.K. and E.L. All authors have read and agreed to the published version of the manuscript.

Funding: The research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration (80NM0018D0004). This research was funded by the NASA ROSES Commercial Smallsat Data Scientific Analysis program (NNH22ZDA001N-CSDSA) as well as the Jet Propulsion Laboratory's Data Science Working Group. Computing resources were leveraged at both the NASA Center for Climate Simulation (NCCS) and the Machine Learning and Affiliated Technologies (MLAT) Lab in the Fowler School of Engineering at Chapman University.

Data Availability Statement: The data have been published and are freely available on Zenodo [72]. The model weights and associated configuration files are also publicly available on HuggingFace [73]. Lastly, the code has been tagged at the time of this paper submission and is publicly available

on GitHub [50]. The Planet radiance data cannot be released publicly due to the terms of the NASA Commercial Smallsat Data Acquisition agreement, but all products generated and models trained have been shared, and all other input data are freely available in the aforementioned Zenodo data store.

Acknowledgments: The authors would like to thank NASA, JPL, NCCS, MLAT Lab., The Spatial Informatics Group, LLC., and the Schmid College of Science and Technology, Chapman University, for supporting this research. The authors would also send thanks to Phil Dennison, from the Geography Department, at the University of Utah, as well as NASA for providing the data, without which this research would not have been possible. Finally, the authors would like to thank the anonymous reviewers for taking the time to read this paper and provide valuable feedback.

Conflicts of Interest: Authors Nicholas LaHaye was employed by the company Spatial Informatics Group LLC. 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. 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. LaHaye, N.; Ott, J.; Garay, M.J.; El-Askary, H.M.; Linstead, E. Multi-Modal Object Tracking and Image Fusion With Unsupervised Deep Learning. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2019**, *12*, 3056–3066. [CrossRef]
2. LaHaye, N.; Garay, M.J.; Bue, B.D.; El-Askary, H.; Linstead, E. A Quantitative Validation of Multi-Modal Image Fusion and Segmentation for Object Detection and Tracking. *Remote Sens.* **2021**, *13*, 2364. [CrossRef]
3. LaHaye, N.; Lee, H.; Easley, A.; Garay, M.J.; Yun, K.; Goodman, A.; Kalashnikova, O.V. A Comparison of Model Complexity, Representative Capabilities, and Performance for Self-Supervised Multi-Sensor Wildifre and SMoke Segmentation and Tracking. In Proceedings of the AGU Fall Meeting 2024. American Geophysical Union, Washington, DC, USA, 9–13 December 2024.
4. Warneke, C.; Schwarz, J.P.; Dibb, J.; Kalashnikova, O.; Frost, G.; Al-Saad, J.; Brown, S.S.; Brewer, W.A.; Soja, A.; Seidel, F.C.; et al. Fire Influence on Regional to Global Environments and Air Quality (FIREX-AQ). *J. Geophys. Res. Atmos.* **2023**, *128*, e2022JD037758. [CrossRef]
5. Williams, A.P.; Abatzoglou, J.T.; Gershunov, A.; Guzman-Morales, J.; Bishop, D.A.; Balch, J.K.; Lettenmaier, D.P. Observed Impacts of Anthropogenic Climate Change on Wildfire in California. *Earth Future* **2019**, *7*, 892–910. [CrossRef]
6. Stavros, N.; Agha, A.; Alkalai, L.; Sirota, A.; Quadrelli, M.; Ebadi, K.; Yun, K. *Smoke Sky: Exploring New Frontiers of Unmanned Aerial Systems for Wildland Fire Science and Applications*; CSIRO: Pasadena, CA, USA, 2019. [CrossRef]
7. Barmpoutis, P.; Papaioannou, P.; Dimitropoulos, K.; Grammalidis, N. A Review on Early Forest Fire Detection Systems Using Optical Remote Sensing. *Sensors* **2020**, *20*, 6442. [CrossRef] [PubMed]
8. 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; pp. 3431–3440. [CrossRef]
9. 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. [CrossRef]
10. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. *arXiv Prepr.* **2021**, arXiv:2105.15203.
11. King, M.D.; Kaufman, Y.J.; Menzel, W.P.; Tanre, D. Remote Sensing of Cloud, Aerosol, and Water Vapor Properties from the Moderate Resolution Imaging Spectrometer (MODIS). *IEEE Trans. Geosci. Remote Sens.* **1992**, *30*, 2–27. [CrossRef]
12. GIBS, N. Worldview Snapshots—wvs.earthdata.nasa.gov. Available online: https://wvs.earthdata.nasa.gov (accessed on 24 December 2024).
13. Knight, E.; Kvaran, G. Landsat-8 Operational Land Imager Design, Characterization and Performance. *Remote Sens.* **2014**, *6*, 10286–10305. [CrossRef]
14. Kanezaki, A. Unsupervised Image Segmentation by Backpropagation. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 1543–1547. [CrossRef]
15. Chen, M.; Artières, T.; Denoyer, L. Unsupervised Object Segmentation by Redrawing. *arXiv* **2019**, arXiv:1905.13539.
16. Aganj, I.; Harisinghani, M.G.; Weissleder, R.; Fischl, B. Unsupervised Medical Image Segmentation Based on the Local Center of Mass. *Sci. Rep.* **2018**, *8*, 13012. [CrossRef]
17. Soares, A.R.; Körting, T.S.; Fonseca, L.M.G.; Neves, A.K. An Unsupervised Segmentation Method For Remote Sensing Imagery Based on Conditional Random Fields. In Proceedings of the 2020 IEEE Latin American GRSS & ISPRS Remote Sensing Conference (LAGIRS), Santiago, Chile, 22–26 March 2020; pp. 1–5. [CrossRef]

18. Zhang, R.; Yu, L.; Tian, S.; Lv, Y. Unsupervised remote sensing image segmentation based on a dual autoencoder. *J. Appl. Remote Sens.* **2019**, *13*, 038501. [CrossRef]
19. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.Y.; et al. Segment Anything. *arXiv* **2023**, arXiv:2304.02643.
20. Thangavel, K.; Spiller, D.; Sabatini, R.; Amici, S.; Sasidharan, S.T.; Fayek, H.; Marzocca, P. Autonomous Satellite Wildfire Detection Using Hyperspectral Imagery and Neural Networks: A Case Study on Australian Wildfire. *Remote Sens.* **2023**, *15*, 720. [CrossRef]
21. Spiller, D.; Carbone, A.; Amici, S.; Thangavel, K.; Sabatini, R.; Laneve, G. Wildfire Detection Using Convolutional Neural Networks and PRISMA Hyperspectral Imagery: A Spatial-Spectral Analysis. *Remote Sens.* **2023**, *15*, 4855. [CrossRef]
22. 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]
23. Real-Time Tracking of Wildfire Boundaries Using Satellite Imagery—Research.Google. Available online: https://research.google/blog/real-time-tracking-of-wildfire-boundaries-using-satellite-imagery/ (accessed on 24 December 2024).
24. Sun, Y.; Jiang, L.; Pan, J.; Sheng, S.; Hao, L. A satellite imagery smoke detection framework based on the Mahalanobis distance for early fire identification and positioning. *Int. J. Appl. Earth Obs. Geoinf.* **2023**, *118*, 103257. [CrossRef]
25. Diner, D.J.; Xu, F.; Garay, M.J.; Martonchik, J.V.; Rheingans, B.E.; Geier, S.; Davis, A.; Hancock, B.R.; Jovanovic, V.M.; Bull, M.A.; et al. The Airborne Multiangle SpectroPolarimetric Imager (AirMSPI): A new tool for aerosol and cloud remote sensing. *Atmos. Meas. Tech.* **2013**, *6*, 2007–2025. [CrossRef]
26. Guerin, D.C.; Fisher, J.; Graham, E.R. The enhanced MODIS airborne simulator hyperspectral imager. In Proceedings of the Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XVII, Orlando, FL, USA, 20 May 2011; Volume 8048, pp. 214–224. [CrossRef]
27. Hook, S.J.; Myers, J.J.; Thome, K.J.; Fitzgerald, M.; Kahle, A.B. The MODIS/ASTER airborne simulator (MASTER)—a new instrument for earth science studies. *Remote Sens. Environ.* **2001**, *76*, 93–102. [CrossRef]
28. Green, R.O.; Eastwood, M.L.; Sarture, C.M.; Chrien, T.G.; Aronsson, M.; Chippendale, B.J.; Faust, J.A.; Pavri, B.E.; Chovit, C.J.; Solis, M.; et al. Imaging Spectroscopy and the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS). *Remote Sens. Environ.* **1998**, *65*, 227–248. [CrossRef]
29. Diner, D. *MISR Experiment Overview*; NASA: Washington, DC, USA, 1999.
30. U.S. Department of Commerce. *Visible Infrared Imaging Radiometer Suite (VIIRS) Sensor Data Record (SDR) User's Guide*; U.S. Department of Commerce: Washington, DC, USA, 2017.
31. DOC; NOAA; NESDIS; NASA. *GOES-R Series Product Definition and Users' Guide*; NASA: Washington, DC, USA, 2019.
32. Planet, T. *Planet Application Program Interface: In Space for Life on Earth*; Planet Development Center: San Francisco, CA, USA, 2018.
33. MacQueen, J.B. Some Methods for Classification and Analysis of MultiVariate Observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*; Cam, L.M.L., Neyman, J., Eds.; University of California Press: Berkeley, CA, USA, 1967; Volume 1, pp. 281–297.
34. Lin, B.; Yu, Z.; Huang, F.; Guo, L. Pool-Based Sequential Active Learning For Regression Based on Incremental Cluster Center Selection. In Proceedings of the 2021 Ninth International Conference on Advanced Cloud and Big Data (CBD), Xi'an, China, 26–27 March 2021; pp. 176–182.
35. Burke, A.; Carroll, M.; Spradlin, C. Finding the Trees in a (Random) Forest: How Do We Get a Representative Sample in a Training Dataset for a Global Land Cover Classification? In Proceedings of the AGU Fall Meeting Abstracts, San Francisco, CA, USA, 11–15 December 2023; Volume 2023, p. IN51C–0429.
36. Grill, J.B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P.H.; Buchatskaya, E.; Doersch, C.; Pires, B.A.; Guo, Z.D.; Azar, M.G.; et al. Bootstrap your own latent: A new approach to self-supervised Learning. *arXiv* **2020**, arXiv:2006.07733.
37. Assran, M.; Duval, Q.; Misra, I.; Bojanowski, P.; Vincent, P.; Rabbat, M.G.; LeCun, Y.; Ballas, N. Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture. *arXiv* **2023**, arXiv:2301.08243.
38. Liao, R.; Kornblith, S.; Ren, M.; Fleet, D.J.; Hinton, G. Gaussian-Bernoulli RBMs Without Tears. *arXiv* **2022**, arXiv:2210.10318.
39. He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; Girshick, R.B. Masked Autoencoders Are Scalable Vision Learners. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 15979–15988.
40. de Boves Harrington, P. Feature expansion by a continuous restricted Boltzmann machine for near-infrared spectrometric calibration. *Anal. Chim. Acta* **2018**, *1010*, 20–28. [CrossRef]
41. Sobczak, S.; Kapela, R. Restricted Boltzmann Machine as Image Pre-processing Method for Deep Neural Classifier. In Proceedings of the 2019 First International Conference on Societal Automation (SA), Krakow, Poland, 4–6 September 2019; pp. 1–5. [CrossRef]
42. Harrington, P.B. Enhanced zippy restricted Boltzmann machine for feature expansion and improved classification of analytical data. *J. Chemom.* **2020**, *34*, e3228. [CrossRef]

43. Ji, X.; Vedaldi, A.; Henriques, J.F. Invariant Information Clustering for Unsupervised Image Classification and Segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9864–9873.
44. Suzuki, S.; Be, K. Topological structural analysis of digitized binary images by border following. *Comput. Vision Graph. Image Process.* **1985**, *30*, 32–46. [CrossRef]
45. Bradski, G. The OpenCV Library. In *Dr. Dobb's Journal of Software Tools*; The University of Utah: Salt Lake, UT, USA, 2000.
46. Wang, Z.; Bovik, A.; Sheikh, H.; Simoncelli, E. Image quality assessment: From error visibility to structural similarity. *Image Process. IEEE Trans.* **2004**, *13*, 600–612. [CrossRef]
47. Wang, Z.; Bovik, A.C. Mean squared error: Love it or leave it? A new look at Signal Fidelity Measures. *IEEE Signal Process. Mag.* **2009**, *26*, 98–117. [CrossRef]
48. Woodhouse, I.H. On 'ground' truth and why we should abandon the term. *J. Appl. Remote Sens.* **2021**, *15*, 041501. [CrossRef]
49. Yang, F.; Du, M.; Hu, X. Evaluating Explanation Without Ground Truth in Interpretable Machine Learning. *arXiv* **2019**, arXiv:1907.06831.
50. LaHaye, N.; Easley, A.; Leet, N. *SIT_FUSE*. 2025. [CrossRef]
51. Harris, C.R.; Millman, K.J.; van der Walt, S.J.; Gommers, R.; Virtanen, P.; Cournapeau, D.; Wieser, E.; Taylor, J.; Berg, S.; Smith, N.J.; et al. Array programming with NumPy. *Nature* **2020**, *585*, 357–362. [CrossRef]
52. Virtanen, P.; Gommers, R.; Oliphant, T.E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods* **2020**, *17*, 261–272. [CrossRef] [PubMed]
53. Dask Development Team. Dask; Dask Development Team. 2016. Available online: https://dask-local.readthedocs.io/en/latest/cite.html (accessed on 10 August 2024).
54. Hoyer, S.; Hamman, J. xarray: N-D labeled arrays and datasets in Python. *J. Open Res. Softw.* **2017**, *5*, 148. [CrossRef]
55. Zenodo. *zarr-developers/zarr-python, Version 3.0.1*; Zenodo/CERN: Geneva, Switzerland, 2025. [CrossRef]
56. Lam, S.K.; Pitrou, A.; Seibert, S. Numba: A LLVM-based Python JIT compiler. In Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC, New York, NY, USA, 15 November 2015; LLVM '15. [CrossRef]
57. Okuta, R.; Unno, Y.; Nishino, D.; Hido, S.; Loomis, C. CuPy: A NumPy-Compatible Library for NVIDIA GPU Calculations. In Proceedings of the Workshop on Machine Learning Systems (LearningSys) in the Thirty-first Annual Conference on Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 4–9 December 2017.
58. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*; Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2019; pp. 8024–8035.
59. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
60. Roder, M.; de Rosa, G.H.; Papa, J.P. Learnergy: Energy-based Machine Learners. *arXiv* **2020**, arXiv:2003.07443.
61. Zenodo. *pytroll/pyresample, Version 1.31.0*; Zenodo/CERN: Geneva, Switzerland, 2024. [CrossRef]
62. Rouault, E.; Warmerdam, F.; Schwehr, K.; Kiselev, A.; Butler, H.; Łoskot, M.; Szekeres, T.; Tourigny, E.; Landa, M.; Miara, I.; et al. *GDAL*; Zenodo/CERN: Geneva, Switzerland, 2025. [CrossRef]
63. Zonca, A.; Singer, L.; Lenz, D.; Reinecke, M.; Rosset, C.; Hivon, E.; Gorski, K. healpy: Equal area pixelization and spherical harmonics transforms for data on the sphere in Python. *J. Open Source Softw.* **2019**, *4*, 1298. [CrossRef]
64. *ssec/polar2grid, Python Package Version 3.2.0*; Zenodo/CERN: Geneva, Switzerland, 2025. [CrossRef]
65. Jordahl, K.; den Bossche, J.V.; Fleischmann, M.; Wasserman, J.; McBride, J.; Gerard, J.; Tratner, J.; Perry, M.; Badaracco, A.G.; Farmer, C.; et al. *geopandas/geopandas, Version 0.8.1*; Zenodo/CERN: Geneva, Switzerland, 2020. [CrossRef]
66. Itseez. Open Source Computer Vision Library. 2015. Available online: https://github.com/itseez/opencv (accessed on 15 January 2025).
67. QGIS Development Team. *QGIS Geographic Information System*; Open Source Geospatial Foundation Project: Beaverton, OR, USA, 2024.
68. Marsocci, V.; Jia, Y.; Bellier, G.L.; Kerekes, D.; Zeng, L.; Hafner, S.; Gerard, S.; Brune, E.; Yadav, R.; Shibli, A.; et al. PANGAEA: A Global and Inclusive Benchmark for Geospatial Foundation Models. *arXiv* **2024**, arXiv:2412.04204.
69. Duderstadt, B.; Helm, H.S.; Priebe, C.E. Comparing Foundation Models using Data Kernels. *arXiv* **2023**, arXiv:2305.05126.
70. Zhang, J.; Lehman, J.; Stanley, K.; Clune, J. OMNI: Open-endedness via Models of human Notions of Interestingness. *arXiv* **2023**, arXiv:2306.01711.
71. Lu, C.; Lu, C.; Lange, R.T.; Foerster, J.; Clune, J.; Ha, D. The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery. *arXiv* **2024**, arXiv:2408.06292.

72. LaHaye, N.; Easley, A.; Lee, H.; Yun, K.; Linstead, E.; Garay, M.; Kalashnikova, O. *Multi-Platform/Multi-Sensor Fire and Smoke Segmentation from the FIREX-AQ 2019 Campaign*; Zenodo/CERN: Geneva, Switzerland, 2025. [CrossRef]
73. LaHaye, N. *FIREX_AQ_2019_Fire_Smoke (Revision e2f8a55)*; HuggingFace: New York, NY, USA, 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.

fire

MDPI

Article

Improving Fire and Smoke Detection with You Only Look Once 11 and Multi-Scale Convolutional Attention

Yuxuan Li [1], Lisha Nie [2], Fangrong Zhou [3], Yun Liu [2], Haoyu Fu [1], Nan Chen [1], Qinling Dai [4] and Leiguang Wang [5,*]

[1] College of Soil and Water Conservation, Southwest Forestry University, Kunming 650224, China; lyx@swfu.edu.cn (Y.L.); yu1463412294@swfu.edu.cn (H.F.); chennan@swfu.edu.cn (N.C.)
[2] College of Forestry, Southwest Forestry University, Kunming 650224, China; nls06@swfu.edu.cn (L.N.); atly9708@swfu.edu.cn (Y.L.)
[3] Joint Laboratory of Power Remote Sensing Technology (Electric Power Research Institute, Yunnan Power Grid Company Ltd., China Southern Power Grid), Kunming 650217, China; zhfr0508@sina.com
[4] Art and Design College, Southwest Forestry University, Kunming 650024, China; daiqinling@swfu.edu.cn
[5] College of Landscape Architecture and Horticulture, Southwest Forestry University, Kunming 650224, China
* Correspondence: leiguangwang@swfu.edu.cn

Abstract: Fires pose significant threats to human safety, health, and property. Traditional methods, with their inefficient use of features, struggle to meet the demands of fire detection. You Only Look Once (YOLO), as an efficient deep learning object detection framework, can rapidly locate and identify fire and smoke objects in visual images. However, research utilizing the latest YOLO11 for fire and smoke detection remains sparse, and addressing the scale variability of fire and smoke objects as well as the practicality of detection models continues to be a research focus. This study first compares YOLO11 with classic models in the YOLO series to analyze its advantages in fire and smoke detection tasks. Then, to tackle the challenges of scale variability and model practicality, we propose a Multi-Scale Convolutional Attention (MSCA) mechanism, integrating it into YOLO11 to create YOLO11s-MSCA. Experimental results show that YOLO11 outperforms other YOLO models by balancing accuracy, speed, and practicality. The YOLO11s-MSCA model performs exceptionally well on the D-Fire dataset, improving overall detection accuracy by 2.6% and smoke recognition accuracy by 2.8%. The model demonstrates a stronger ability to identify small fire and smoke objects. Although challenges remain in handling occluded targets and complex backgrounds, the model exhibits strong robustness and generalization capabilities, maintaining efficient detection performance in complicated environments.

Keywords: fire and smoke object detection; YOLO11; wildfire; attention mechanism

1. Introduction

Wildfires are a natural component of forest ecosystems and play a crucial role in maintaining ecological balance and biodiversity. Moderate wildfires can promote forest species renewal, release stored nutrients, and reintroduce them into the soil cycle, thereby nourishing the forest ecosystem [1,2]. However, improper wildfire management can lead to severe consequences, such as large-scale tree mortality, destruction of forest structures, and loss of habitats for flora and fauna. Additionally, wildfires release significant amounts of carbon dioxide and other pollutants, exacerbating global climate change [3,4]. Against the backdrop of climate change, the frequency and intensity of wildfires are increasing annually, further threatening the stability of forest ecosystems. Therefore, timely monitoring and rapid response to fires and their resulting smoke not only help prevent the spread of fires and protect forest resources but also mitigate the damage caused by climate change to

ecosystems. Moreover, the occurrence of wildfires is closely linked to human activities, as humans are one of the primary causes of wildfires, whether through direct ignition or indirect trigger [5]. Thus, fire prevention not only reduces the ecological damage caused by fires but also safeguards human lives and property from threats [6], making it essential to enhance fire prevention and control capabilities. To address this challenge, efforts are being made to explore more efficient technologies and methods for preventing and extinguishing wildfires, thereby protecting the ecological environment and promoting the sustainable development of human society.

Over the years, traditional fire detection methods have been widely used for wildfire monitoring, including manual patrols, watchtower observations, and physical sensor detection. Considering personnel safety, the monitoring range of manual patrols is relatively limited, making it difficult to detect fires in a timely manner [7]. The fire response capability of watchtowers is also constrained by their field of view, and their construction is limited by terrain and budget, resulting in less than ideal monitoring effectiveness [8]. As a result, an increasing number of people are adopting various sensor technologies to enhance fire monitoring capabilities, such as smoke sensors and temperature sensors [9–11]. While these sensor technologies are cost-effective and easy to implement, they are more suitable for indoor environments, require manual intervention to verify the authenticity of alarms, and cannot provide specific fire information, such as the exact location, scale, and intensity of the fire [12,13]. Additionally, temperature triggers require reaching a certain threshold to respond, and the process of smoke particles reaching the sensors can be influenced by natural factors such as wind speed [14], leading to delays that ultimately reduce detection accuracy.

With the continuous advancement of surveillance technology, visual fire detection technology has become an effective alternative to traditional fire detection methods. When smoke or fire appears within the camera's field of view, the system can immediately detect the fire, demonstrating significant potential for early fire detection. Additionally, visual fire detection can provide critical information about fire development, such as the spread speed and direction, which is essential for assessing fire evolution and minimizing losses [15,16]. Currently, vision-based fire detection technologies are primarily divided into machine learning methods and deep learning methods. Machine learning methods [17] rely on manually designed features to identify targets, with the drawback of requiring feature selection and parameter tuning, and the selected features may not be sufficiently representative. To address these limitations, deep learning technology has become the preferred choice in the field of fire detection due to its ability to automatically extract features, achieve high detection accuracy, and exhibit strong adaptability [18,19]. Object detection technology in deep learning has garnered widespread attention for its capability to automatically identify and locate fire and smoke objects in images or video streams. Object detection tasks are generally divided into two categories: two-stage detection and one-stage detection. Two-stage detection first generates candidate regions that may contain objects; in the second stage, these candidate regions are classified and subjected to bounding box regression to determine the target's location and category. Representative algorithms include R-CNN [20], Fast R-CNN [21], and Faster R-CNN [22], among others. In contrast, one-stage detection methods directly predict the object's location and category from the input image, completing both classification and regression tasks in a single forward pass. Representative algorithms include SSD [23] and the YOLO (You Only Look Once) series [24,25].

Although two-stage detection methods offer higher accuracy, their process of first generating candidate regions and then performing classification and regression results in greater computational overhead and longer processing times. Therefore, in real-time wildfire detection tasks that demand both speed and accuracy, the one-stage YOLO ob-

ject detection framework has been widely improved and applied [26,27]. To verify the effectiveness of YOLOv3 for real-time forest fire detection on small drones, Zhentian Jiao et al. [28] proposed an improved algorithm based on YOLOv3-tiny and tested it on a drone platform. The experiments demonstrated that this method met expectations in terms of both detection speed and accuracy. For rapid detection of ship fires in complex and dynamic marine environments, Huafeng Wu et al. [29] constructed a high-quality dataset, introduced multi-scale detection techniques, and incorporated the SE (Squeeze and Excitation) attention mechanism. This effectively enhanced the accuracy and efficiency of small object detection. Additionally, they accelerated model convergence using transfer learning. The experimental results showed that the detection speed fully met real-time requirements. To address the challenge of accurately identifying different types of forest fires in complex backgrounds, Qilin Xue et al. [30] proposed an improved YOLOv5-based model called FCDM. By introducing the SIoU loss function, the CBAM attention mechanism, and an enhanced BiFPN feature fusion network, the model performed exceptionally well in fire classification and detection tasks. It demonstrated significant potential for practical applications. In the future, researchers plan to explore model lightweighting to improve deployment efficiency. To tackle the difficulty of detecting small smoke targets in complex backgrounds, Xin Chen et al. [31] proposed the RepVGG-YOLOv7 model. This model incorporated the ECA attention mechanism, the SIoU loss function, the RepVGG structure, and an improved non-maximum suppression algorithm. These innovations significantly boosted feature extraction capabilities and improved inference efficiency through lossless compression techniques. Experimental results indicated that the model achieved a precision of 95.1% in complex backgrounds and small target detection. Future work will focus on further optimization for edge devices and exploring the integration of multimodal data. Additionally, to address the issues of high false alarm and missed detection rates in existing fire detection technologies, Tianxin Zhang et al. [32] proposed an improved YOLOv8-based fire detection algorithm, YOLOv8-FEP. By adding a large object detection head, introducing the EMA attention mechanism, and designing a PAN-Bag structure, the model excelled in detecting large-scale fires and smoke object and effectively improved robustness and accuracy in complex scenarios. Experiments indicated that YOLOv8-FEP achieved higher precision, lower false alarm rates, and lower missed detection rates in fire and smoke detection tasks, making it more suitable for real-time detection. Future studies will focus on expanding the training dataset and optimizing the algorithm to enhance the model's performance in special environments. Chenmeng Zhao et al. [33] proposed a lightweight deep learning model called SF-YOLO (Smoke and Fire-YOLO), which is based on an improved version of YOLOv11. It incorporates the C3k2 module with a dual-path residual attention mechanism and the SEAMHead detection head with an embedded attention mechanism, along with a dynamically adjusted W-SIoU loss function. These innovations significantly enhance the model's ability to detect targets in situations with occlusion, blurred boundaries, and complex backgrounds. Experiments conducted on the self-built dataset S-Firedata and the public dataset M^4SFWD show that SF-YOLO outperforms existing models in terms of detection accuracy (with mAP50 and mAP50-95 improving by 2.5% and 2.4%, respectively) as well as lightweight design. Hongying Liu et al. [34] proposed a Transformer-based multi-scale feature fusion network called TFNet. This network enhances feature extraction through the SRModule, utilizes the CG-MSFF Encoder to fuse multi-scale features, and employs the WIOU loss function to optimize localization accuracy. Experimental results demonstrate that TFNet performs excellently on the D-Fire and M4SFWD public datasets, outperforming existing advanced models in metrics such as precision, recall, F1 score, mAP50, and mAP50–95. At the same time, it maintains a good balance in terms of parameter size and computational complexity,

making it suitable for real-time applications. From the past to the present, from CNN to Transformer, researchers have been working on improving YOLO from different angles to make it more suitable for fire and smoke object detection task requirements. We use Table 1 to more directly observe what problems the researchers have solved and how to solve the remaining problems in the future.

Table 1. Summary of the relevant studies.

Approach	Key Contributions	Future Work
ESFD-YOLOv8 [26]	Develop a real-time system for early and accurate smoke and fire detection in complex environments.	Address data imbalance, misdetection in complex scenes, and real-time optimization.
EA-YOL [27]	Solve issues of low precision, small and dense target detection, sample imbalance, and balancing real-time performance with accuracy in existing fire detection models.	Tackle challenges in small target detection, data scarcity, and complex background interference.
Using UAV and YOLOv3 [28]	Provide an efficient, low-cost UAV solution for resource-constrained field environments.	Improve detection of small fires in forests.
I-YOLOv4-tiny + S [29]	Overcome the accuracy limitations and computational complexity of traditional fire detection methods and deep learning models in complex maritime environments.	Future work will focus on optimizing model lightweighting, improving accuracy, incorporating contextual information, and expanding applications.
FCD [30]	Address challenges in distinguishing forest fire types, improving detection accuracy, small target detection, real-time performance, and dataset limitations.	Future work will optimize model performance, integrate multiple detection models, design lightweight models, and expand datasets.
RepVGG-YOLOv7 [31]	Solve the problem of low smoke detection accuracy in complex backgrounds and small targets, while balancing model complexity with detection speed.	Future plans include expanding datasets and reducing model computation and parameters.
YOLOv8-FE [32]	Overcome the low detection accuracy of traditional methods in complex environments prone to false alarms and missed detections due to background interference.	Future work aims to enhance detection speed, meet higher real-time requirements, and reduce model complexity for resource-constrained devices.
SF-YOLO (Smoke and Fire-YOLO) [33]	Solve the insufficient detection accuracy of traditional methods in complex environments.	Future plans include incorporating environmental covariates into model training and exploring multispectral data for smoke and flame detection.
TFNet [34]	Effectively extract features of small targets and sparse smoke, maintaining high accuracy in complex backgrounds while reducing model parameters and computational complexity.	Future plans include expanding dataset scale and exploring more lightweight network architectures.

It is evident that many efforts and studies have been conducted using the YOLO framework to address issues in visual fire detection, such as the diverse types, colors, and shapes of smoke and fire, the complexity of the environment, and the practicality of model deployment [15]. These challenges remain key research topics in visual fire detection [35]. The latest YOLO series, YOLO11, was released by the Ultralytics team at the end of September 2024. YOLO11 utilizes an improved backbone and neck architecture to enhance feature extraction, offering a faster processing speed while maintaining the best balance between accuracy and performance. It can be seamlessly deployed on edge devices,

cloud platforms, and GPU-equipped systems, ensuring maximum flexibility [36]. Whether the latest YOLO11 can better address the challenges in fire and smoke object detection tasks or what its advantages are remains to be seen. This study conducts experiments to discuss the performance of YOLO11 in fire and smoke object detection tasks.

This study addresses the following specific problems and conducts the following work:

1. The YOLO series has consistently played an active role in fire and smoke target detection. This study first explores the advantages of the YOLO11 network in fire and smoke target detection tasks. The result show that YOLO11 achieves the best balance between detection accuracy, detection speed, and model complexity for fire and smoke object detection tasks. Moreover, all YOLO models demonstrate higher detection accuracy for smoke than for fire, and the YOLO series can detect fires at an earlier stage. YOLO11 will be an important tool for our future fire and smoke object detection research.
2. To address the variability of smoke and fire objects across different scales and environments, this study introduces the Multi-Scale Convolutional Attention (MSCA) module to enhance YOLO11 (for simplicity, the improved YOLO11 is referred to as YOLO11s-MSCA in this paper). The components of MSCA are characterized by high efficiency, low computational cost, and minimal parameters, ensuring that the model's detection speed and algorithm complexity do not significantly increase.
3. Observations from the results of YOLO11s-MSCA on the D-Fire show that YOLO11s-MSCA improves the model's detection accuracy without noticeably compromising detection speed or increasing model complexity. Compared to YOLO11s, the improved model detects small targets in images more accurately and demonstrates a stronger understanding of environmental contexts. However, it still struggles with identifying the occluded object, and the main sources of errors remain the complex environmental variations.
4. The generalization performance of our YOLO classic model and YOLO11s-MSCA model is verified through the overall metrics and visual details on the Fire and Smoke Dataset and CBM-Fire. Even when the dataset is changed, the models maintain high accuracy and recall, confirming that the conclusions drawn from the D-Fire dataset are applicable. At the same time, different datasets of the same model are compared, and the conclusion is drawn that the construction strategy of the dataset (data volume, scene richness, classification of interference items) has a significant impact on the performance and stability of the detection model. Especially when the interference terms are clearly labeled as independent categories, the robustness and accuracy of the model can be significantly improved in complex environments.

2. Materials and Methods

2.1. Datasets

To improve the accuracy of training results and enhance the generalization ability of deep convolutional neural network models, it is crucial to select datasets that can simulate real fire scenarios. This research comprehensively evaluates the model's performance by assessing its results on the D-Fire dataset, the Fire and Smoke Dataset, and CBM-Fire.

The D-Fire [37] fire and smoke object detection dataset contains 21,527 images, with 17,221 used for training the model and 4306 for testing. The dataset includes 1164 images featuring only fire cases, 5867 images containing only smoke cases, and 4658 images where both fire and smoke are present. These extensive annotated data provide a wealth of samples for the model. To improve the model's robustness, the dataset also contains 9838 images without fire or smoke, including instances often misidentified as fire and smoke, such as fog and bright light. To enhance the model's generalization ability and simulate

complex scenarios, the dataset's sources are highly diverse, covering scenes from the internet, schools, technology parks, and parks—all with intricate backgrounds. Moreover, the dataset distinguishes between day and night images, and it includes various camera interference factors, such as raindrops and spiders, which reflect real-world complexities. The number of instances in the dataset is detailed in Table 2, ensuring that the model can be trained on a substantial number of samples. Typical examples from the dataset are illustrated in Figure 1, showcasing the diversity of its sources.

Table 2. The number of bounding boxes per category in the D-Fire dataset.

Category	Training Set	Test Set
Fire	14,692	2869
Smoke	11,865	2307

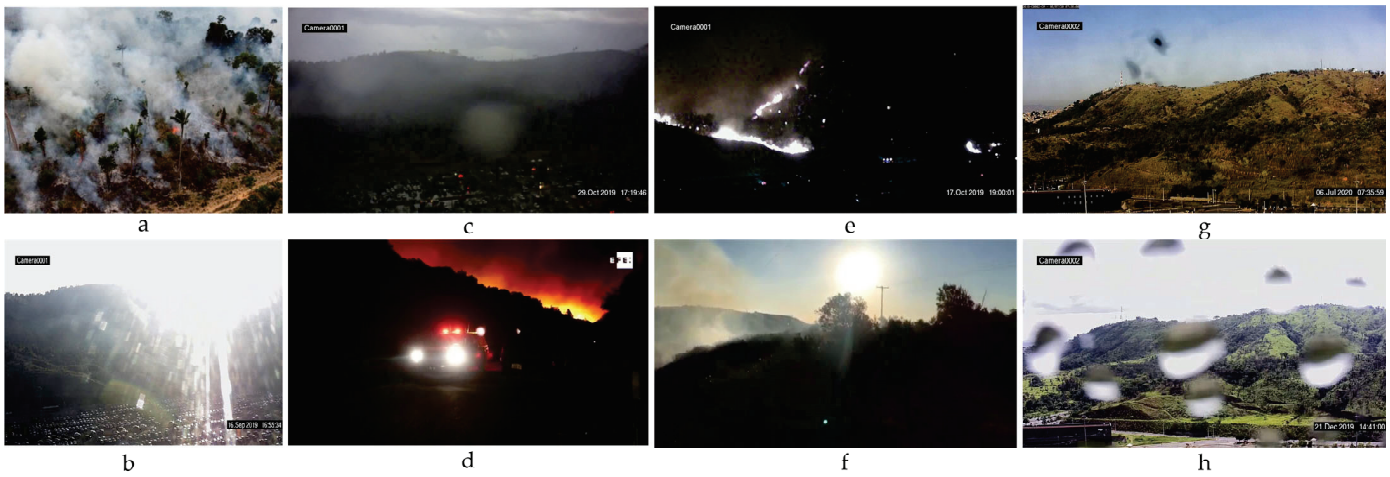


Figure 1. Examples from the D-Fire dataset: (**a**) top–down view, (**b**) high-brightness non-fire target, (**c**) foggy weather, (**d**) nighttime fire and car headlights, (**e**) nighttime fire, (**f**) bright morning sun, (**g**) lens insect, and (**h**) lens water droplets.

The Fire and Smoke Dataset [38] is the second dataset used in this study, primarily to validate the model's generalization ability and result applicability. This publicly available dataset consists of two parts, from which we downloaded FireSmokeDataset_part1. It contains 12,813 training images, 6068 validation images, and 2237 test images—sufficient data volume to effectively support model training and validation. The dataset covers a wide range of scenarios, including indoor and outdoor environments, urban areas, industrial sites, and forests. It also includes both low- and high-resolution images, as well as multi-instance cases. These diverse scenes simulate various real-world complexities, ensuring the model undergoes comprehensive training and testing even under challenging conditions. The dataset is divided into three categories: smoke, fire, and other classes. The "other classes" include images that are easily mistaken for fire, such as sunrise and sunset, clouds, streetlights, campfire embers, reflections, vehicle headlights, emergency lights of various vehicles, and lights from public electronic screens. The inclusion of these categories enhances the model's ability to recognize interfering factors, thereby improving its robustness in practical applications. The dataset is sourced from a variety of origins, including other public datasets, images downloaded from the internet, and frames extracted from videos, encompassing images from daytime to nighttime and under different weather conditions. Typical examples from the dataset are illustrated in Figure 2, effectively simulating complex real-world scenarios.

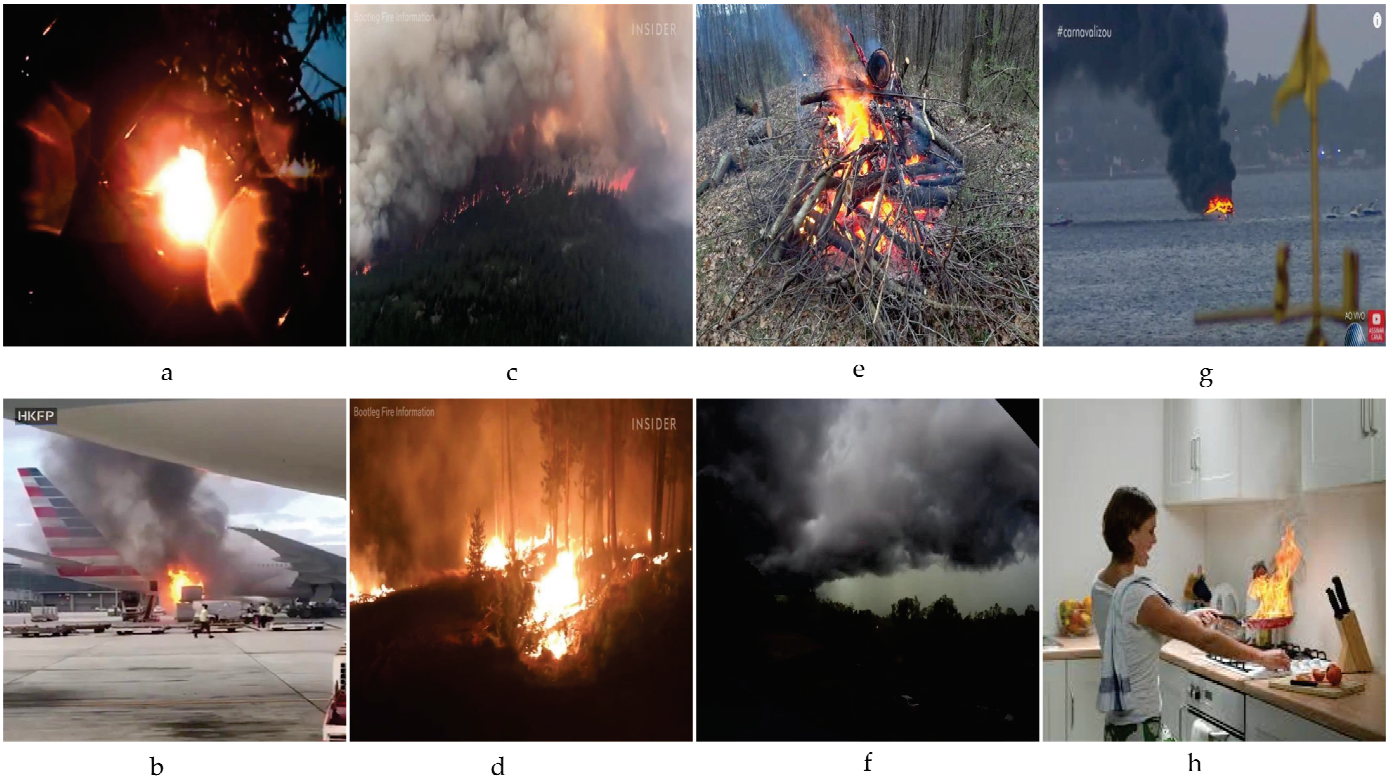


Figure 2. Examples from the Fire and Smoke Dataset ((**a**) low-resolution highlighting target; (**b**) airport fire; (**c**) drone platform image; (**d**) high-resolution fire at night; (**e**) near fire; (**f**) cloudy day smoke target; (**g**) sea boat fire; and (**h**) indoor fire).

To address the issue of insufficient quantity and low quality in fire target detection datasets, Xin Geng et al. proposed a new fire detection dataset called CBM-Fire [39]. This dataset contains 2000 images of flames and smoke, divided into two categories—flame and smoke—at a ratio of 2:1. The images cover a wide range of fire scenarios and are annotated using LabelMe software. The dataset is split into training, testing, and validation sets in an 8:1:1 ratio.

When users download the dataset, the images are already organized by category. By simply reading the corresponding .txt files with the provided code, the images can be sorted into separate folders accordingly, as shown in Table 3 (CBM-Fire dataset partitioning ratio). The dataset features high diversity, precise annotations, and excellent quality, making it well-suited for the training and evaluation of fire detection algorithms. The related resources have been publicly released on the open-source repository at: https://github.com/GengHan-123/yolov9-cbm (accessed on 15 April 2025).

Table 3. CBM-Fire dataset partitioning ratio.

CBM-Fire	Training Set	Test Set	Validation Set
2000	1620	200	180

2.2. Model Overview

2.2.1. Proposed Method

The latest YOLO network framework consists of four main components: the input, backbone, neck, and head, as illustrated in parts a, b, and c of Figure 3. The input layer processes the image and applies data augmentation [40]. The backbone network is responsible for extracting multi-scale features from the input image. The neck uses a Feature

Pyramid Network (FPN) [41] to aggregate features from different scales, enhancing the model's ability to detect objects of various sizes and improving multi-scale semantic representation. The detection head consists of several convolutional layers designed to produce detection results.

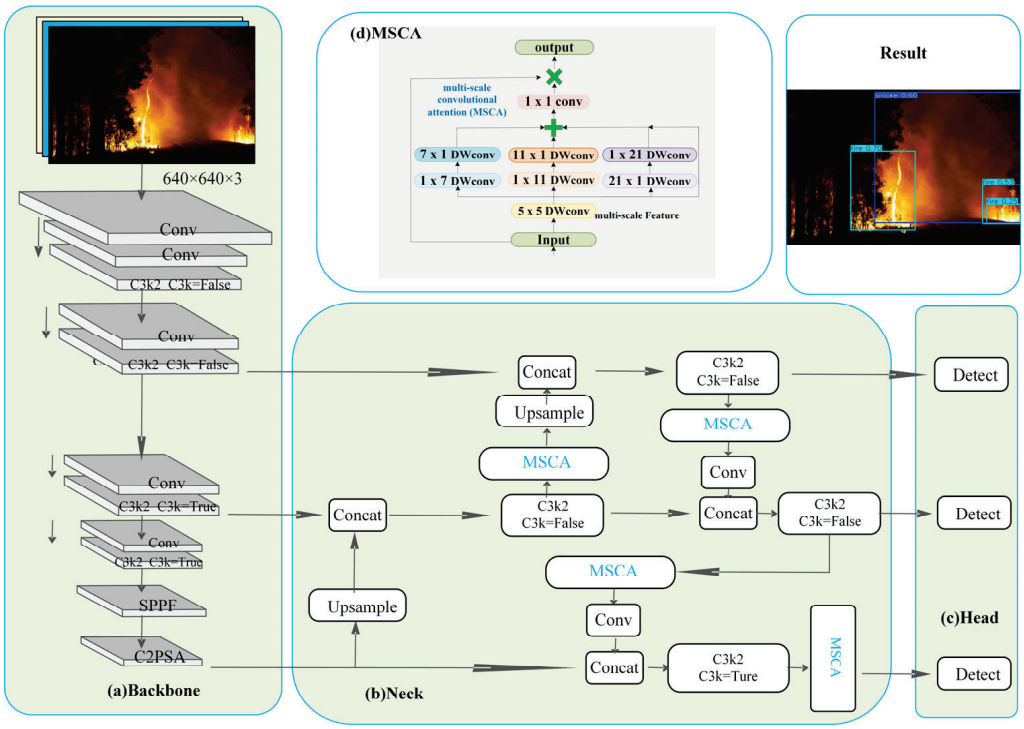


Figure 3. YOLO11s-MSCA framework ((**a**) YOLO11 backbone; (**b**) YOLO11 neck; (**c**) YOLO11 head; and (**d**) MSCA module).

This paper focuses on YOLO11, a model developed by the Ultralytics team, similar to YOLOv8, with several improvements based on YOLOv8 [42]. Key information can be found on the website https://docs.ultralytics.com/models/yolo11/ (accessed on 15 April 2025). The backbone modules of YOLO11 include the Conv composite module (comprising Conv2d, batch normalization [43], and the SiLU activation function [44]), the C3k2 module, the Spatial Pyramid Fast Pooling (SPPF) module [26], and the Cross Stage Partial Spatial Attention (C2PSA) module. C3k2 and C2PSA are novel modules introduced by YOLO11. The C3k2 module enhances the model's adaptability to complex scenes by employing parallel convolution and optimized parameter settings, effectively handling multi-scale features. The module achieves efficient feature extraction through parallel convolution branch and parameter configurable design. Its core adopts a two-way structure: one way directly transfers shallow features, and the other way processes deep features through an optional C3k (variable convolution kernel) or standard bottleneck module and finally concatenates and fuses. When C3k is false, C3k2 is C2f, and when C3k is true, the module is C3k2. The details are shown in Figure 4. Meanwhile, the C2PSA module combines the Cross Stage Partial (CSP) [45] structure and the Pyramid Squeeze Attention (PSA) [46] mechanism, further boosting feature extraction capabilities, particularly excelling in multi-scale feature extraction. CSP will process the feature map by dividing it into two parts: one part is directly transmitted, while the other part is processed through the PSA module. Finally, the two parts are concatenated and fused. By emphasizing spatial correlations in feature maps, C2PSA ensures that the model focuses on crucial regions, thereby striking a balance between computational cost and detection accuracy (refer to Figure 5 for details). In

the detection head, YOLO11 produces detection results at three different scales by utilizing feature maps generated from the backbone and neck.

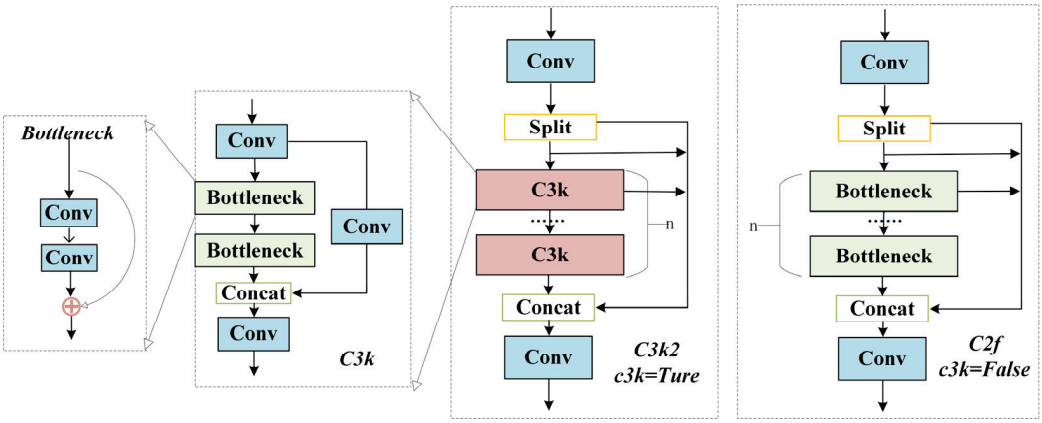


Figure 4. C3k2 module.

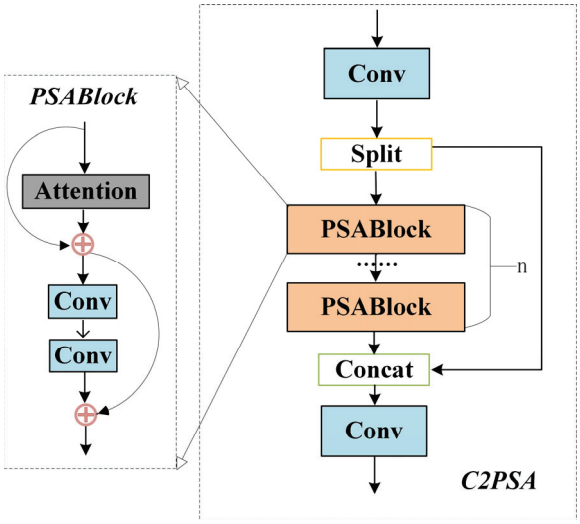


Figure 5. C2PSA module.

The improvements in this study focus on the neck of the YOLO11 network. The neck network adopts a structure like YOLOv8, integrating the Feature Pyramid Network (FPN) and Path Aggregation Network (PANet) [47]. By introducing top–down pathways and lateral connections, it effectively fuses high-level semantic information with low-level spatial details, enhancing the model's ability to detect objects of varying sizes [48]. This structure efficiently processes the multi-scale features extracted from the backbone network. In YOLO11, the C3k2 module is introduced into the neck to enable the model to extract features at different scales and facilitate multi-scale information interaction. This allows the model to capture more complex spatial features, which are then passed to the detection head. Increasing the number of neurons in the neck network may further optimize the advantages of feature fusion. After the C3k2 module enables multi-scale feature interaction, the MSCA module is introduced to focus on important regions' multi-scale contextual information, highlighted in blue in Figure 3.

To maintain a balance between accuracy and model complexity, this study currently adopts a single-module approach, ensuring high precision without significantly increasing computational cost.

2.2.2. Multi-Scale Convolutional Attention (MSCA)

The MSCA (Multi-Scale Convolutional Attention) mechanism was proposed by Meng-Hao Guo et al. [49] to address the challenge of efficiently encoding contextual information in semantic segmentation tasks. It leverages multi-scale convolutional features to activate spatial attention [50], thereby encoding contextual information more effectively. To tackle the variability of smoke and fire types, as well as the complexity of the environment, this study introduces MSCA into the YOLO11 framework to focus on the multi-scale and contextual information of objects. The key components of MSCA are as follows:

Depth-wise Convolution: used to aggregate local information. Unlike standard convolutions that operate across the entire input feature map, depth-wise convolutions perform convolution independently on each input channel [51]. The kernel depth matches the number of input channels, while the kernel width and height remain relatively small. The primary advantage of this convolution method lies in its computational efficiency, as it avoids convolution operations between different channels, significantly reducing the number of parameters and calculations.

Multi-branch Depth-wise Strip Convolutions: used to capture multi-scale contextual information. Each branch employs strip convolutions, a lightweight convolution method that approximates a standard k × k convolution using a pair of k × 1 and 1 × k convolutions [52]. Compared to standard convolutions, strip convolutions significantly reduce computational cost and the number of parameters while maintaining performance. This design balances efficiency and accuracy, enabling the model to process spatial features across multiple scales effectively.

The 1 × 1 Convolution: used to model relationships between different channels [53] and generate attention weights. It reduces the number of feature maps while maintaining a relatively low parameter count.

The high efficiency, low computational cost, and minimal parameters of each component in the MSCA module are key reasons for incorporating it into the fire and smoke detection task. It ensures that accuracy is improved without significantly increasing model complexity or reducing detection speed.

Specifically, the computational process of MSCA is as shown in Equations (1) and (2):

$$\mathrm{MSCA} = \mathrm{Conv}_{1\times1}\left(\sum_{i=0}^{3} \mathrm{Scale}_i(\mathrm{DW\_Conv}(In))\right) \tag{1}$$

$$\mathrm{Out} = \mathrm{MSCA} \otimes In \tag{2}$$

Here, ***In*** represents the input features, **MSCA** and **Out** denote the attention map and output, respectively, ⊗ signifies the element-wise matrix multiplication operation, **DW_Conv** stands for depth-wise convolution, $\mathbf{Scale_i}$ represents the i-th scale branch, and $\mathbf{Scale_0}$ is the identity connection. A detailed diagram of the MSCA module is shown in Figure 6.

By adopting this approach, MSCA effectively captures multi-scale features and generates adaptive attention weights, enhancing the model's performance. This research integrates this module into the YOLO11s model, which significantly boosts the model's accuracy. The specific results will be discussed in the next section. MSCA is a plug-and-play module, and we do not need to enter additional parameters. It is an added single module, and no changes are made to the other modules of the source code.

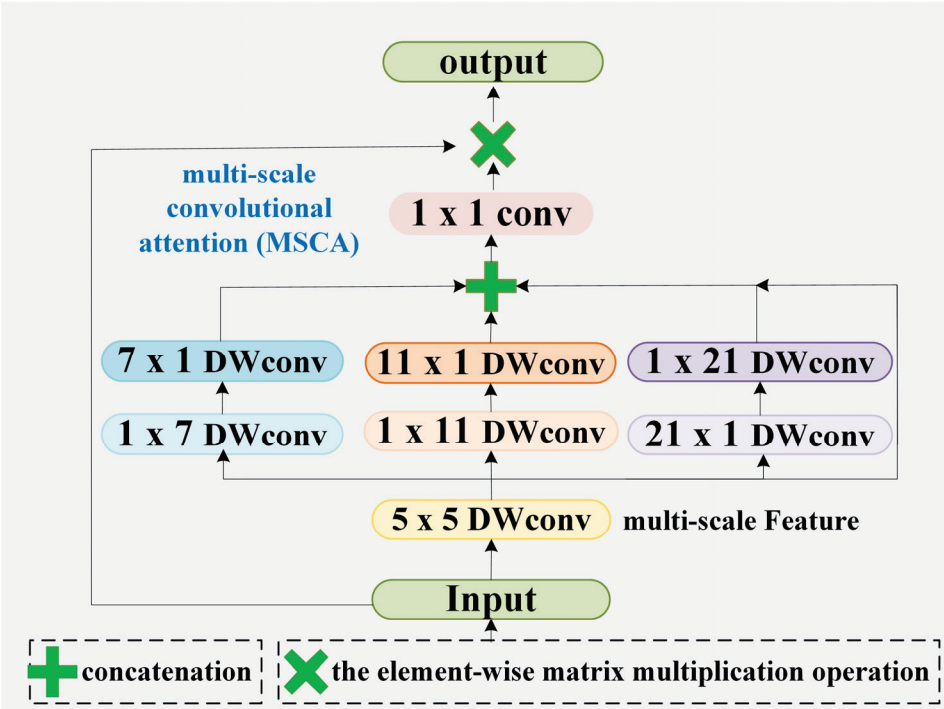


Figure 6. MSCA module (where DWConv represents depth-wise convolution, and k × k indicates the convolution kernel size).

3. Experiments and Analyses

3.1. Evaluation Metrics Section

To improve the model, evaluate its performance, and compare differences between models, selecting appropriate evaluation metrics is crucial. This study employs multiple metrics related to object detection [37,54]. Given the severe danger posed by wildfires, timely and accurate detection of fires and their accompanying smoke is of utmost importance. Therefore, the accuracy of fire and smoke object detection is the primary focus. Additionally, the number of missed detections serve as a critical basis for assessing model performance and guiding future improvements. To evaluate the model's performance in terms of high precision, this paper employs metrics such as precision, recall, mAP50, and mAP50-95. Furthermore, to meet real-time requirements and evaluate the model's computational efficiency, this study uses FPS (frames per second) and GFLOPS (giga floating-point operations per second) as metrics. The specific formulas are provided in Equations (3)–(6).

1. **Precision**

Precision measures the proportion of correctly identified targets among all the targets detected by the model. The higher the precision, the fewer false positives the model has.

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3}$$

2. **Recall**

Recall measures the proportion of correctly detected objects among all actual objects. It represents the ratio of correctly detected objects among all objects of interest (each smoke and fire object). A higher recall indicates fewer missed detections by the model.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4}$$

Here, TP (True Positive) refers to targets correctly identified and localized by the model, meaning the bounding box detected by the model fully matches the ground truth box and is classified as the positive class (target class). FP (False Positive) refers to cases where the model incorrectly detects background regions or non-target regions as targets, meaning the model predicts the existence of a target that does not actually exist. FN (False Negative) refers to real targets that the model fails to detect, meaning actual targets are neither recognized nor localized by the model.

3. **Mean Average Precision (mAP)**

The area under the precision–recall curve for a single class is called average precision (AP), and the mean of AP across all classes is referred to as mean average precision (mAP) [55]. mAP50 refers to the average precision calculated at an Intersection over Union (IoU) threshold of 0.5. mAP50-95 is an extension of mAP, which calculates the average precision across multiple IoU thresholds ranging from 0.5 to 0.95 with a step size of 0.05. This value reflects the model's capability in bounding box regression.

$$\mathrm{mAP} = \frac{1}{m}\sum_{i=1}^{m} \mathrm{AP}(i), \tag{5}$$

$$\mathrm{AP} = \int_0^1 P(r)\mathrm{d}r, \tag{6}$$

4. **GFLOPS**

GFLOPS is one of the metrics used to measure the computational complexity of a model. It reflects the number of floating-point operations required for the model to perform forward inference. The higher the GFLOPS value, the more floating-point operations the model needs to perform, which increases the computational demands and may slow down the inference speed, especially when hardware resources are limited.

5. **FPS**

FPS is a critical metric for evaluating the real-time performance of a model, particularly in applications such as video processing, real-time object detection, and augmented reality [56]. FPS represents the number of images or video frames processed by the model per unit of time, directly reflecting its ability to meet real-time requirements.

3.2. Experimental Configuration and Parameter Setting

To ensure that performance differences stem from the model's design and functionality rather than external factors such as hardware configuration or hyperparameter settings, it is essential to meticulously document and control these external variables.

The deep learning training platform used in this study has powerful computing capabilities and efficient parallel processing abilities, making it suitable for training and inference on large-scale datasets. The system configuration includes the Windows Server 2022 Standard 64-bit operating system, equipped with two Intel(R) Xeon(R) Gold 6330 processors, with a base frequency of 2.00 GHz and a maximum frequency of 3.10 GHz, and 256 GB of high-speed memory, providing ample memory support for handling large datasets and complex models. The system is also equipped with a Tesla T4 GPU (15206 MiB), supporting the CUDA computing platform, which accelerates model training and inference processes in deep learning tasks, significantly reducing training time. The deep learning framework used is PyTorch 2.4.1, with CUDA version 11.8, the Ultralytics 8.3.18 toolkit, and Python version 3.10.15, ensuring flexibility and efficiency in the development environment.

The training process consists of 100 epochs, with an input image size of 640 × 640 pixels. The batch size is set to 20, and the optimization algorithm used is Stochastic Gradient

Descent (SGD), with all other hyperparameters set to their default values. The initial learning rate (lr0) is set to 0.01, and the SGD momentum factor is set to 0.937 by default. Additionally, the number of worker threads for data loading is set to 0 to ensure efficient memory usage. Mosaic data augmentation is disabled in the final 10th epoch of training. The evaluation metrics in the following sections are calculated based on these hyperparameter settings to ensure comparison and analysis. The key hyperparameter settings are shown in Table 4.

Table 4. Hyperparameter settings.

Parameters	Values
imgsz	640
epochs	100
batch	20
close_mosaic	10
optimizer	SGD
lr0	0.01
momentum	0.937

3.3. Experimental Results and Discussion Analysis

3.3.1. Comparative Analysis of the Advantages of YOLO11

The main reason for the generation-by-generation turnover of the YOLO model is to continuously improve its performance, speed, accuracy, and adaptability, thereby meeting the growing demands of computer vision tasks. In order to explore how YOLO11 is superior in the direction of fire and smoke target detection on the D-Fire dataset, this research compares YOLO11 with previous classical methods of the YOLO family. All our experiments and index calculations are under the same device and the same hyperparameter settings, ensuring that the index difference is caused by the model improvement part, as described in Section 3.2. This experiment selects models that are representative and can be run on our equipment.

To meet the needs of more diverse computer vision applications, YOLOv5 aims to improve the accuracy and ease of use of real-time object detection while optimizing resource usage. By improving the network architecture and training strategy, the detection accuracy is improved to meet the demands of complex scenarios. In addition, computational and memory resource consumption is optimized to ensure efficient operation on resource-constrained devices and expand the types of inspection task [57]. Table 5 shows that in this task, the YOLOv5s is 1.1% less refined overall compared to the YOLO11s, and all other metrics are worse than the YOLO11s except for the FPS improvement of 32.3.

YOLOv6 is dedicated to solving the key problems of speed and accuracy balance, model deployment utility, quantization performance degradation and label assignment, and loss function fitness in the field of object detection. It redesigns the network architecture and training strategy for the diverse needs of speed and accuracy in different scenarios and strives to achieve the best balance between the two [58]. As seen in Table 5, YOLOv6s has larger GFLOPS as well as higher FPS, which differ from YOLO11s by 22.7 and 20.8, respectively. However, compared to YOLO11s, the overall accuracy loss is 1.7%, and the mAP50 for the fire target is 2.2% lower, which is the largest difference among all the metrics, and all the other metrics are lower than YOLO11s.

YOLOv7 aims to address the balance between speed and accuracy in real-time object detection, while focusing on improving the detection of small objects. It further reduces model complexity and computational resource consumption and improves training efficiency by optimizing network architecture and training strategies. In addition, YOLOv7 enhances the generalization capability of the model, enabling it to better adapt to diverse

application scenarios and meet different task requirements [59]. As shown in Table 5, the overall precision of YOLOv7 is only 69%, a difference of 8.1% compared to YOLO11, mainly since the detection precision of YOLOv7 for fire object is only 62.7%, which is significantly lower than that of the YOLO11 algorithm.

Table 5. Comparison of YOLO11 with the classic YOLO method (D-Fire).

Model	Class	Precision (%)	Recall (%)	mAP50 (%)	mAP50-95 (%)	GFLOPS	FPS
YOLOv5s	All	76.0	69.6	76.7	44.8	23.8	344.8
	Smoke	80.3	76.3	82.6	51.8		
	Fire	71.7	62.8	70.8	37.8		
YOLOv6s	All	75.4	69.9	76.0	44.3	44.0	333.3
	Smoke	80.0	77.3	82.8	52.0		
	Fire	70.9	62.5	69.2	36.6		
YOLOv7	All	69.0	**73.1**	75.2	39.9	103.2	117.6
	Smoke	75.3	74.7	78.4	43.7		
	Fire	62.7	71.6	71.9	36.2		
YOLOv8s	All	77.2	70.4	77.5	45.6	28.4	322.6
	Smoke	81.4	77.6	83.7	52.9		
	Fire	73.0	63.1	71.3	38.3		
YOLOv9s	All	**77.7**	72.1	**78.3**	**45.9**	26.7	250.0
	Smoke	82.5	79.3	84.4	53.3		
	Fire	72.9	64.9	72.1	38.5		
YOLOv10s	All	76.5	69.3	76.3	44.8	24.4	312.5
	Smoke	79.8	77.6	82.8	52.2		
	Fire	73.3	61.0	69.8	69.8		
YOLO11n	All	75.3	67.4	74.6	42.8	**6.3**	**434.8**
	Smoke	79.9	74.2	80.5	49.7		
	Fire	70.8	60.5	68.6	35.9		
YOLO11s	All	77.1	70.1	77.4	45.5	21.3	312.5
	Smoke	81.7	77.3	83.3	52.7		
	Fire	72.4	63.0	71.4	38.2		

The goal of YOLOv8 is to further improve the performance and applicability of the YOLO series in real-time object detection tasks through technological innovations and architectural optimizations, while maintaining its ease of use and high efficiency, with few innovations and bias towards engineering practice, which can be effectively scaled up between different hardware configurations [60]. As shown in Table 5, YOLO11s has 7.1 fewer GFLOPS compared to YOLOv8s, with other metrics slightly reduced and an overall accuracy decrease of 0.1%. Although the performance difference is minimal, YOLO11s exhibits lower model complexity.

YOLOv9 addresses the problem of data loss during transmission through deep networks by introducing innovative methods such as Programmable Gradient Information (PGI) and Generalized Efficient Layer Aggregation Network (GELAN), a problem that can lead to biased gradient streams, which in turn affects the training effectiveness and prediction accuracy of the model [61]. As shown in Table 5, compared with YOLOv9s, the overall accuracy of YOLO11s decreased by 0.6%, GFLOPS decreased by 5.4, and FPS increased by 62.5. YOLO11s shows a slight decrease in accuracy, but it significantly improves model complexity and reduces detection speed.

The YOLOv10 research team aims to further advance the performance efficiency boundaries of YOLO in terms of both post-processing and model architecture and comprehensively optimizes each component of YOLO from the perspectives of efficiency and accuracy, greatly reducing computational overhead and improving performance [62]. As

shown in Table 5, compared with YOLOv10s, the overall accuracy of YOLO11s is improved by 0.6%, the accuracy of smoke detection is improved by 1.9%, other accuracy indicators are also improved, and GFLOPS reduced by 3.3.

YOLO11 provides a version with five parameters of n, s, m, l, and x gradually increasing. Although the version size is different, it can still achieve a good balance between accuracy and speed. In order to achieve the balance of accuracy, real-time performance, and model complexity under our equipment, this experiment compares the differences between YOLO11n and YOLO11s. Other versions will surge in the number of parameters, which does not meet our low complexity requirements and real-time performance when detecting fire and smoke objects [36]. It can be seen from Table 5, compared with YOLO11 s, the YOLO11 n parameter is smaller, the GFLOPS is 6.3, which is only 30% of YOLO11 s, and the FPS increases by 112.3, which meets the requirements of real-time detection and lower model complexity, but the accuracy is 1.8% lower than the overall accuracy of YOLO11s. YOLO11n provides us with a new choice based on specific task requirements. Due to the great harm of fire in the fire and smoke detection task, this study puts the accuracy in the first consideration. The improved baseline model in this paper is YOLO11s.

The variation in evaluation metrics during model training is shown in Figure 7. It can be observed that YOLOv7 performs the worst in terms of precision, while the differences among the other models are not significant. For recall, YOLO11n performs the worst, and YOLOv7 achieves the highest score. YOLOv9s and YOLO11s exhibit the most stable changes, indicating that these two models have a high precision rate and can stably and continuously observe targets. YOLO11n and YOLOv7 show the worst mAP results, while YOLOv9s performs the best. The performance of YOLO11s and several other models under different IoU thresholds varies little. Although YOLO11s does not achieve the best results in all metrics, each metric remains stable and ranks highly, demonstrating its advantage of stability and balance.

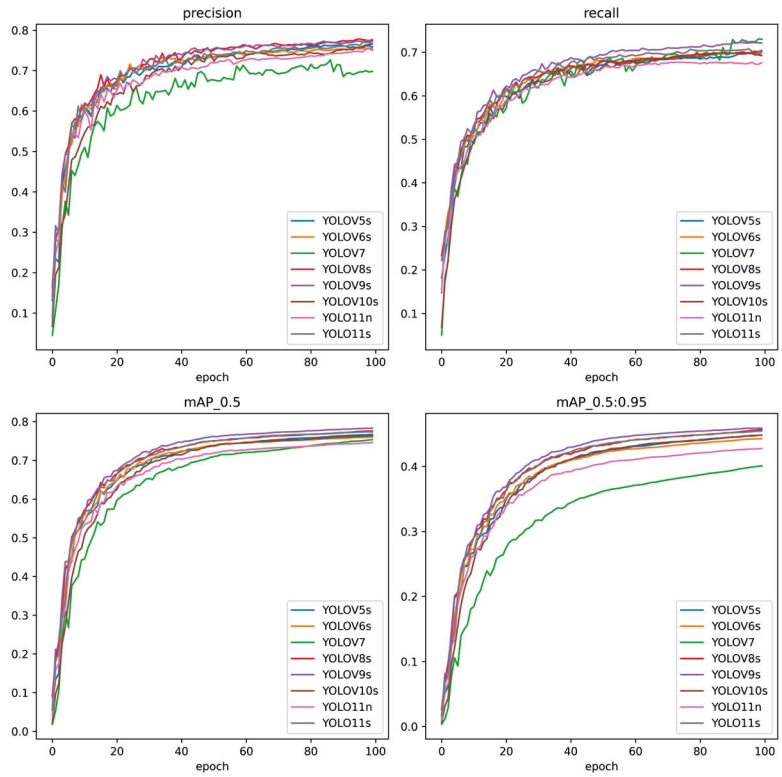


Figure 7. Model training evaluation metrics variation. YOLO series baseline model training evaluation metrics variation.

Through comparative analysis, it is evident that all YOLO models exhibit higher detection accuracies for smoke objects compared to fire objects, indicating that YOLO is effective for early fire detection. Although each model has its strengths and weaknesses, the differences in accuracy are not significant. YOLO11 demonstrates a clear advantage in fire and smoke object detection tasks, outperforming other methods in the YOLO family in terms of balancing accuracy, speed, and real-time performance. This makes YOLO11 another important tool for future fire and smoke detection tasks. Additionally, the superior performance of other models in detecting fire or smoke can provide valuable insights for future improvements.

3.3.2. YOLO11s-MSCA Model Performance Analysis

The previous section verified that the YOLO11 model demonstrates significant advantages in fire and smoke detection tasks. Since the performance of deep learning models is closely related to specific tasks, selecting appropriate improvement methods is crucial. In this section, our study focuses on enhancing the YOLO11 model for fire and smoke detection, making it more suitable for these tasks. Based on the characteristics of fire and smoke summarized by previous studies [19], this study introduces the MSCA module to strengthen YOLO11, with the specific framework shown in Figure 3. We start comparing the effects of our improvements from the model training stage.

In the process of YOLO series model training and verification, the optimal model is determined by minimizing the bounding box loss (box_loss), confidence loss (dfl_loss), and classification loss (cls_loss). As seen in Figure 8, YOLO11n, YOLO11s, and YOLO11s-MSCA losses are gradually minimized during the training process, and the YOLO11s-MSCA model has the best minimization effect, the model converges faster, and YOLO11n is the worst; in the validation set, there is no significant difference in the localization loss of the boundaries of YOLO11s and YOLO11s-MSCA. YOLO11n has the worst localization ability, the three models have little difference in confidence loss, and YOLO11s-MSCA accelerates the fastest convergence in classification loss, suggesting that the introduced attentional mechanism can focus on our region of interest faster.

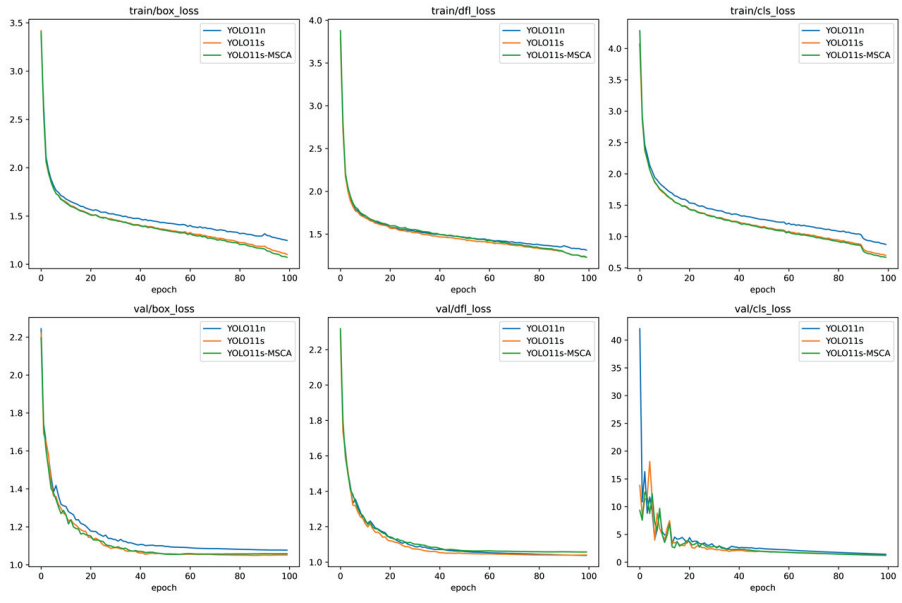


Figure 8. YOLO model training loss variation.

Due to different models, RT-DETR loss function combines classification loss, L1loss and IOUloss to optimize the boundary box regression effect. The decrease in Giou_loss

indicates that the boundary frame positioning changes, and the decrease in L1_loss indicates that the coordinate prediction is more stable. Figure 9 shows that during model training, boundary positioning gradually tends to be stable, and classification loss, although it has a downward trend, has been fluctuating violently. In the validation process, the positioning ability of the RT-DETR model tends to stabilize faster than that of YOLO series models, but the classification loss (cls_loss) tends to stabilize more slowly than that of YOLO series models.

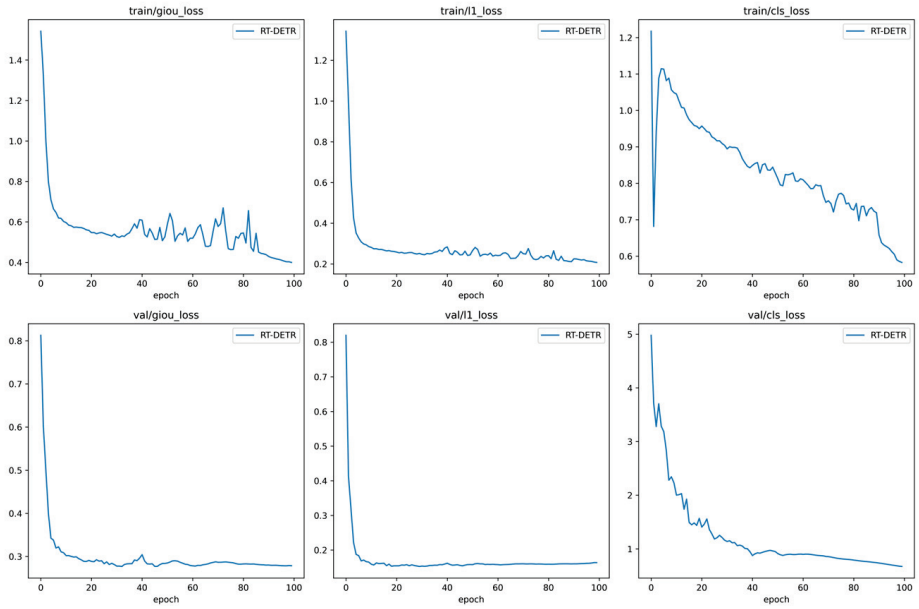


Figure 9. RT-DETR model training loss variation.

Figure 10 shows the comparison of accuracy curves between YOLO11s-MSCA and other classical target detection models in the training process. All indexes of the RT-DETR model are the worst. The Precision_Curve in Figure 10 shows that YOLO11S-MSCA has the highest accuracy rate and no significant differences with YOLO11s in other indicators, indicating that the introduced module mainly improves the classification ability, but the model boundary regression has no significant differences, which is in line with the reality. We have not improved the model boundary regression. Future improvements can be made in these directions. After that, we made a comparative analysis of the model from the evaluation indicators. Since YOLO11n had the worst effect, we did not make the comparison in Table 6.

Table 6. Evaluation accuracy of YOLO11s-MSCA validation set (D-Fire).

Model	Class	Precision (%)	Recall (%)	mAP50 (%)	mAP50-95 (%)	GFLOPS	FPS
RT-DETR	All	71.5	63.2	70.2	39.5	103.4	68.0
	Smoke	72.1	71.2	74.9	45.2		
	Fire	71.0	55.2	65.5	33.8		
YOLO11s	All	77.1	**70.1**	77.4	45.5	**21.3**	**312.5**
	Smoke	81.7	77.3	83.3	52.7		
	Fire	72.4	63.0	71.4	38.2		
YOLO11s-MSCA	All	**79.7**	69.8	**77.9**	**45.8**	22.5	222.2
	Smoke	84.5	76.9	84.0	53.3		
	Fire	74.9	62.6	71.8	38.2		

To enhance the comprehensiveness of the comparison, we compare with RT-DETR, the first Transformer architecture model capable of end-to-end detection in real-time scenarios. Experimental results of RT-DETR on the COCO dataset indicate that it significantly outperforms popular YOLO series real-time detectors in both speed and accuracy [63]. However, in our D-Fire dataset, the RT-DETR model only achieves an overall accuracy of 71.5%, with other metrics also showing significant differences when compared to YOLO. Most importantly, the model's GFLOPS reaches 103.4, which is 4.6 times that of YOLO11s-MSCA, while the detection data consist of only 68 frames, which is one-third of YOLO11s-MSCA. Therefore, for fire and smoke object detection tasks, YOLO11s-MSCA outperforms RT-DETR. The specific results can be found in Table 6.

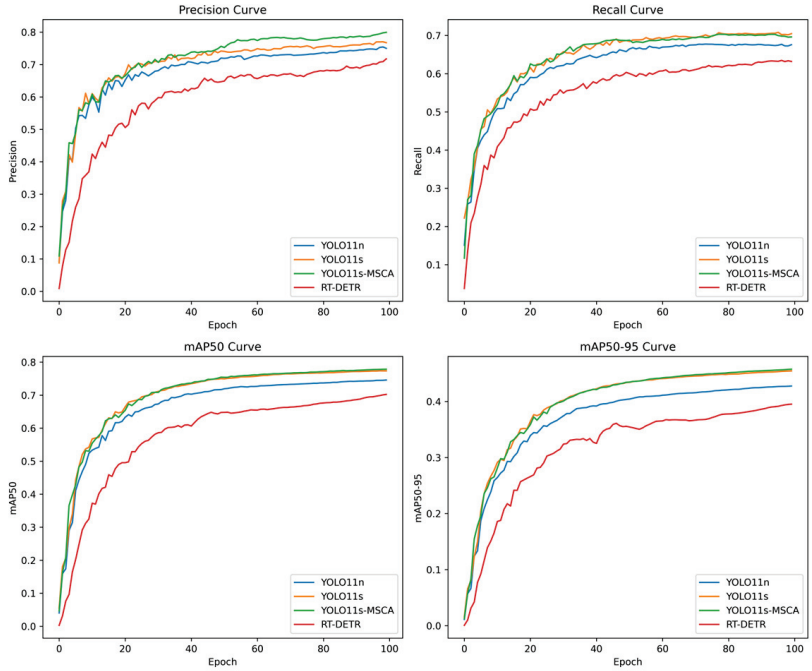


Figure 10. Comparison of training evaluation metrics between YOLO11s-MSCA and classical models. As can be seen from Table 6, the overall accuracy rate of the improved algorithm has increased by 2.6%, and the contribution to the accuracy rate is mainly reflected in the recognition of smoke, which has increased by 2.8%, and the recognition of fire, which has increased by 2.5%, indicating that the introduction of the MSCA module can pay more attention to the smoke object, while other indicators have not been greatly improved. The improved overall precision reaches 79.7% with a recall of 69.8%, indicating that the model has fewer false and missed detections. With an mAP50 of 77.9%, the model still has good target detection ability in complex scenes, and with an mAP50-95 of 45.8%, the model still has objective detection accuracy in a larger IOU (Intersection over Union) range [54]. The increase in GFLOPS for the improved model only adds 1.2, mainly because the modules introduced use deep convolution, multi-scale banded convolution, and 1×1 convolution, which are computationally efficient with low parameters. With an FPS of 222.2, the improved model does not significantly lose real-time detection performance or increase model complexity while improving accuracy.

Accurate identification of fire events is exceptionally important, the accuracy of smoke and fire detection should be in the first place, because the hazards of fire occurrence are very large, and the improved model meets the requirements of high accuracy. The occurrence of fire is accompanied by the occurrence of smoke; because the diffusion of smoke in the early stage of fire is much higher than that of fire, the improvement of the model of the more accurate identification of the smoke target allows the relevant personnel to monitor the target area in a timely manner, which is exceptionally important for the early detection

of fire. The improved model did not significantly increase the model complexity. Our improvements fulfill the requirements for fire and smoke object detection.

While metrics analysis can provide a quantitative assessment for the object detection model and help researchers understand the overall performance of the model, visual analysis provides more detailed information about the model and can help us gain a deeper understanding of the model's behavior under different conditions.

By analyzing the visual results, it can be seen that YOLO11s is accurate in classification and localization in smoke and fire detection tasks, but there are still problems in many details, and through the observation of the results of the verification set, the improved YOLO11s-MSCA model in this paper is effective, and the advantages of the algorithm improvement are as follows:

1. As shown in Figure 11b,e, the improved method in this paper is more accurate in the detection of small objects. The small object fires that are not labeled in the label are identified accurately, which proves the effectiveness of the multi-scale spatial attention introduced in this paper. The model focuses on objects of different scales (smoke and fire objects in this paper).
2. The YOLO11s-MSCA model is more accurate in locating the target frame. For example, the a in Figure 11 correctly locates the range of the smoke object and does not identify the scattering of the light as smoke. In c, the cloud of smoke object and background are correctly localized. Spatial attention makes the model focus on the correct important areas.
3. The YOLO11s-MSCA model can more accurately identify the differences between fire objects and highlighted fire-like objects and can correctly classify them as in Figure 11a,c,d. The improved model in Figure 11f can correctly distinguish smoke targets at night. The advantage comes from the ability to understand the context. Smoke is a fire companion organism, and the two basically exist at the same time.

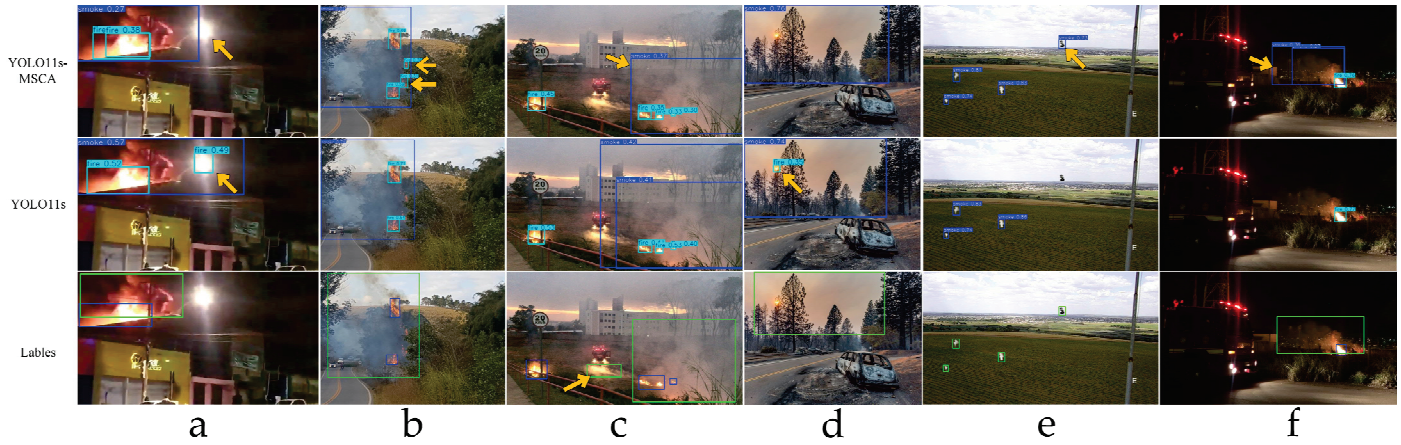


Figure 11. Analysis of typical examples of visual advantages in YOLO11s-MSCA (D-Fire). (**a**) accurately locate the fire and lights at night. (**b**) accurately identify small targets. (**c**) accurately locate the smoke target and distinguish between light and fire. (**d**) accurate sunset recognition. (**e**) accurate distant target recognition. (**f**) Accurate smoke detection near flames at night.

The improving limitations of model detection results are as follows:

1. The introduced contextual spatial attention module focuses on contextual information, which also causes some loss of accuracy in our results, e.g., our model does not detect small object white smoke in the white wall background. Small object fire in the background of a smoke, as in Figure 12g, identifies the highlight as fire, and smoke in the background of a grayish white, as in Figure 12e, does not detect smoke object.

2. For occluded objects, the algorithm does not recognize them correctly, as in Figure 12e,f, occluded targets are not recognized.
3. Although the improved algorithm has enhanced the detection performance for small targets, challenges remain, as illustrated in Figure 12c,d. This limitation is primarily attributed to the complexity of the environment.
4. The YOLO11s-MSCA model still has the problem of misidentification for feature similarities, e.g., in Figure 12a, smoke in the distance is similar to clouds, scattered light similar to smoke (e.g., b), and sunset in red clouds (e.g., i).
5. The complexity of the environment and the different performance of smoke and fire in different environments are the most important sources of error rate. The performance of smoke and fire comes from the difference in combustion material, the change in light, the change in day and night, the distance from the video sensor and so on. The processing of this variability is still an important research direction for us in the future. As in Figure 12h, the recognition accuracy of smoke and fire objects decreases at night.

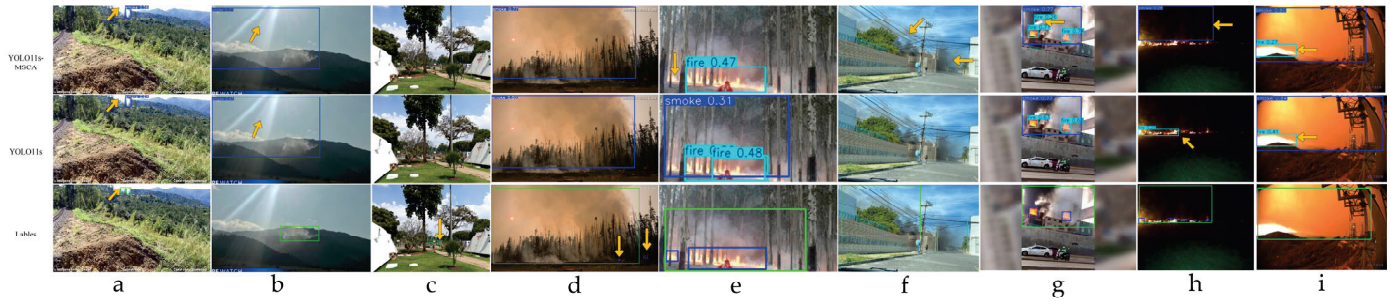


Figure 12. Analysis of typical examples of visual limitations in YOLO11s-MSCA (D-Fire). (**a**). white smoke on sky background. (**b**). scattered light resembling smoke in color. (**c**). blurred small objects at distance. (**d**). tiny objects. (**e**). extensive smoke coverage and partially obscured fires. (**f**). smoke objects partially obscured by electrical wires. (**g**). highlighted regions around flames. (**h**). nighttime thin smoke and small fires. (**i**). Sunset at the horizon's edge.

3.3.3. Comparison of Model Generalization Experiments

In this study, in order to verify that the model has good generalization ability, robustness, and practical application adaptability, the YOLO11s-MSCA model was systematically verified on two representative datasets: Fire and Smoke and CBM-Fire. Detailed training parameter settings are described in Section 3.2 of this article.

As can be seen through the bolded font in Table 7, YOLOv9 was the most accurate model during the validation of the Fire and Smoke Dataset, but YOLO11 and its improved version were still the more balanced models across all metrics. The overall precision, recall, and mAP of YOLO11s-MSCA are higher than those of YOLO11s by 0.3%, 0.8%, and 0.6%, respectively, although there is an increase of 1.2 and a decrease of 33 frames in GFLOPS and FPS metrics, respectively, and the balance of improving accuracy for fire and smoke object detection tasks without affecting the other performances. Our improvements are effective and generalizable. Future research on YOLO11 is an important direction for our fire and smoke object detection mission. Compared to RT-DETR, which successfully introduces the Transformer architecture into the field of real-time object detection, YOLO11s-MSCA has an overall accuracy that is 0.5 lower. However, all other metrics are better than those of RT-DETR. Notably, in terms of the balance between accuracy, speed, and model complexity, YOLO11s-MSCA is the best, making it more suitable for our fire and smoke object detection domain.

Table 7. Comparison of evaluation metrics between YOLO11s-MSCA and classical YOLO methods (Fire and Smoke Dataset).

Model	Class	Precision (%)	Recall (%)	mAP50 (%)	mAP50-95 (%)	GFLOPS	FPS
YOLOv9s	All	78.7	72.8	**79.2**	**50.4**	26.7	51.3
	Fire	83.0	84.0	89.3	58.8		
	Other	69.0	56.1	61.6	33.8		
	Smoke	84.1	78.3	86.6	58.5		
YOLOv10s	All	77.5	72.2	77.7	49.0	24.5	131.6
	Fire	81.4	83.0	88.2	58.1		
	Other	68.6	55.2	59.0	31.5		
	Smoke	82.4	78.3	86.0	57.3		
RT-DETR	All	**79.0**	71.2	76.5	46.9	103.4	64.9
	Fire	83.7	83.3	87.7	56.4		
	Other	69.0	54.0	58.2	30.4		
	Smoke	84.2	76.2	83.5	54.0		
YOLO11s	All	78.3	72.4	78.6	49.5	**21.3**	**142.9**
	Fire	83.1	83.9	89.4	58.6		
	Other	67.8	54.5	31.9	31.9		
	Smoke	84.0	78.8	87.0	58.0		
YOLO11s-MSCA	All	78.5	**73.2**	**79.2**	50.3	22.5	109.9
	Fire	83.0	84.0	89.4	59.0		
	Other	68.1	56.2	60.7	33.4		
	Smoke	84.4	79.5	87.5	58.6		

The overall performance of the YOLO11s and YOLO11s-MSCA models is understood through metrics, and the details of the model's performance on the Fire and Smoke Dataset are as follows:

1. YOLO11s and YOLO11s-MSCA still perform well in fire and smoke object detection tasks, accurately classifying and localizing most objects in an image. Figure 13i correctly distinguishes fire objects from lights; j correctly localizes smoke objects by eliminating trees from the smoke background.
2. In this dataset, YOLO11s-MSCA is still able to focus more on small objects in the model than the YOLO11s model and classify them correctly, e.g., it can focus on the candles in the pictures in Figure 13a,i, differentiating between the lights in the background of the fire and the small object flames in b, f, and k. The introduced Multi-Scale Convolutional Attention mechanism still focuses on the multi-scale information of the object in the image.
3. The YOLO11s-MSCA model recognizes the sun in clouds, as shown in Figure 13d, and correctly distinguishes smoke objects in snowy landscapes, as in Figure 13c,e, and thin smoke in h. These are not small objects, their correct distinction comes from the model's understanding of the object and the object's surroundings. The Multi-Scale Convolutional Attention mechanism this research introduces can still focus on the object's contextual information.
4. For occluded objects, the model still fails to recognize them, as in Figure 13g. Similarly, our introduction of the MSCA module causes some small objects similar to the background to be misdistinguished, and the most important errors still originate from the variability of the environment and the diversity of fire and smoke.

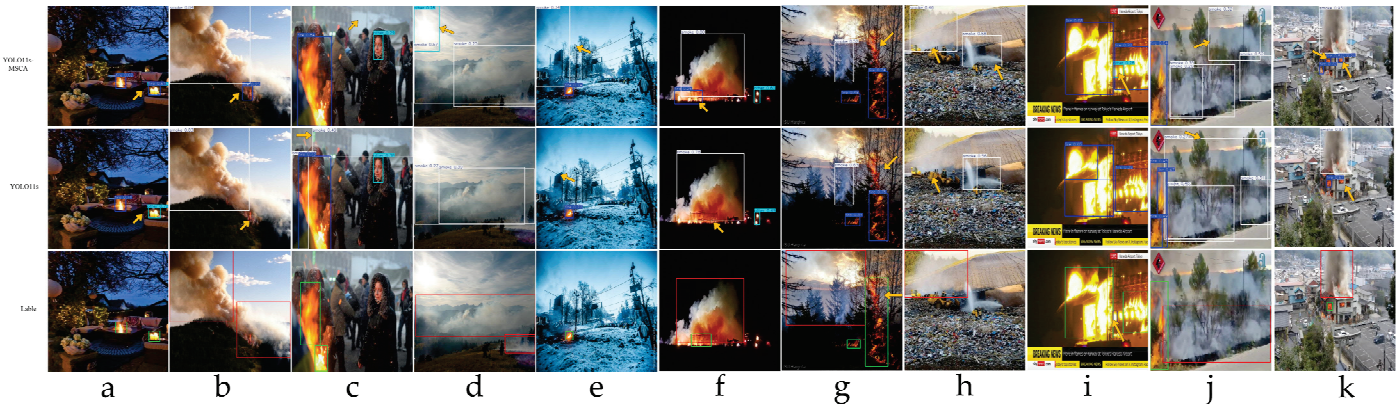


Figure 13. Analysis of typical examples of visual effects in YOLO11s-MSCA (Fire and Smoke Dataset). (**a**). candles and small fires. (**b**). small fire objects. (**c**). robust snow-smoke discrimination. (**d**). correctly identify the sun. (**e**). smoke recognition under snow interference. (**f**). more accurate positioning. (**g**). occluded target recognition inaccuracy. (**h**). near-field thin smoke. (**i**). light sources in fire environments. (**j**). precision smoke target localization. (**k**). small-target fire.

YOLO11s is validated to have high precision, high recall, high detection speed, and low model complexity in fire and smoke target detection tasks, and while not the best in all metrics, it ranks among the top in every metric and has the best balance in all metrics. At the same time, our improved model YOLO11s-MSCA is effective. The introduced module improves the recognition ability of the model to different sizes of targets in the image and the understanding of context information, so that the overall accuracy has been significantly improved. This model has more accurate identification and localization of small object fires and smoke, ability to identify thin smoke around fires, smoke objects in snow, etc. The problem summarization and improvement of the improved model on the D-Fire dataset hold true on the second dataset as well.

The model's generalization capability on unseen data has become a critical metric for evaluating its practical value. Building upon the CBM-Fire dataset, this study continues to investigate the generalization performance of YOLO11s-MSCA through comparative experiments, aiming to explore the model's adaptability in fire and smoke object detection scenarios.

As presented in Table 8, YOLOv9s achieves an overall accuracy of 82.3%, which is 2.7% lower than that of YOLO11s-MSCA (85.0%). While the difference in FPS between the two models is 23.9, the remaining evaluation metrics exhibit more pronounced disparities. YOLOv10s underperforms YOLO11s-MSCA across all metrics, with an overall accuracy of just 77.7%, and particularly substantial gaps in recall and mAP50, registering differences of 11.4% and 11.1%, respectively. The RT-DETR model demonstrates the weakest performance, with an accuracy of only 57.4%, and significantly inferior results across all other metrics. Notably, its GFLOPS is 4.6 times higher than that of YOLO11s-MSCA, indicating considerably greater computational cost.

Among all evaluated models, the best-performing metrics (highlighted in bold) are consistently observed in either YOLO11s or YOLO11s-MSCA. YOLO11s-MSCA, in particular, achieves the highest overall accuracy while maintaining a balanced trade-off among other key indicators such as recall, mAP50-95, GFLOPS, and FPS. Importantly, the model attains a 3% gain in accuracy at the cost of only a 1.2-point increase in GFLOPS, underscoring the efficiency and effectiveness of the improvements. These enhancements prove to be robust across all three datasets, further affirming the generalizability of YOLO11s-MSCA in fire and smoke object detection tasks.

Table 8. Comparison of YOLO11 and classical object detection methods (CBM-Fire).

Model	Class	Precision (%)	Recall (%)	mAP50 (%)	mAP50-95 (%)	GFLOPS	FPS
YOLOv9s	All	82.3	64.4	76.5	46.9	26.7	175.4
	Fire	74.8	65.4	74.9	40.9		
	Smoke	89.7	63.4	78.1	52.9		
YOLOv10s	All	77.7	59.0	70.8	46.2	24.4	131.6
	Fire	75.0	61.5	69.4	41.7		
	Smoke	80.5	56.5	72.2	50.7		
RT-DETR	All	57.4	48.9	49.5	27.3	103.4	69.9
	Fire	63.1	60.1	63.4	33.5		
	Smoke	51.8	37.7	35.7	21.0		
YOLO11s	All	82.0	**71.3**	80.6	**53.6**	**21.3**	**212.8**
	Fire	78.8	75.6	81.2	81.2		
	Smoke	85.3	67.0	80.1	59.2		
YOLO11s-MSCA	All	**85.0**	70.4	**81.9**	52.0	22.5	151.5
	Fire	80.1	75.5	84.4	48.1		
	Smoke	90.0	65.2	79.4	55.9		

Additionally, as illustrated in Figure 14, the model's behavior under the CBM-Fire dataset—augmented with image rotations—is analyzed in detail. The performance of YOLO11s-MSCA is summarized as follows:

1. **Enhanced Confidence and Localization Accuracy:** The YOLO11s-MSCA model demonstrates superior confidence scores and more precise localization capabilities. As shown in Figure 14a–c (highlighted with yellow arrows), the model accurately detects fire and smoke targets. Notably, in Figure 14b, it correctly identifies a fire instance that is absent in the ground truth annotations, and in Figure 14c, it effectively localizes a broader region of smoke. These results confirm the efficacy of the introduced attention mechanism in improving target recognition.
2. **Improved Detection of Small Objects:** YOLO11s-MSCA exhibits enhanced sensitivity to small-scale targets, as demonstrated in Figure 14a,d, where it successfully detects small fire sources (again indicated by yellow arrows). In contrast to YOLO11s, which misclassifies a firefighter's red uniform as fire in Figure 14a, YOLO11s-MSCA correctly identifies the actual fire target. This further supports the effectiveness of the attention mechanism in directing focus towards semantically relevant regions.
3. **Robustness Across Diverse Environments:** Comparative experiments across three distinct datasets—each differing in source and environmental conditions—demonstrate that YOLO11s-MSCA maintains robust performance in the face of varying complexity and scene dynamics.
4. **Limitations in Background-Dominant Scenes:** As shown in Figure 14c, although YOLO11s-MSCA captures a larger smoke area, it fails to identify smoke against a sky background. Similarly, in Figure 14e, where the entire image is engulfed in smoke, the model is unable to distinguish the smoke target. This indicates that while the MSCA module enhances environmental understanding, it may also reduce the model's discriminability when the target blends seamlessly with the background.
5. **Challenges with Occlusion and Visual Similarity:** In certain scenarios, such as in Figure 14f, the model struggles with occluded targets. However, the primary cause of detection errors remains the high variability of the environment. For instance, in Figure 14f, the fire color closely resembles the yellowish reflection on nearby leaves, leading to misclassification.

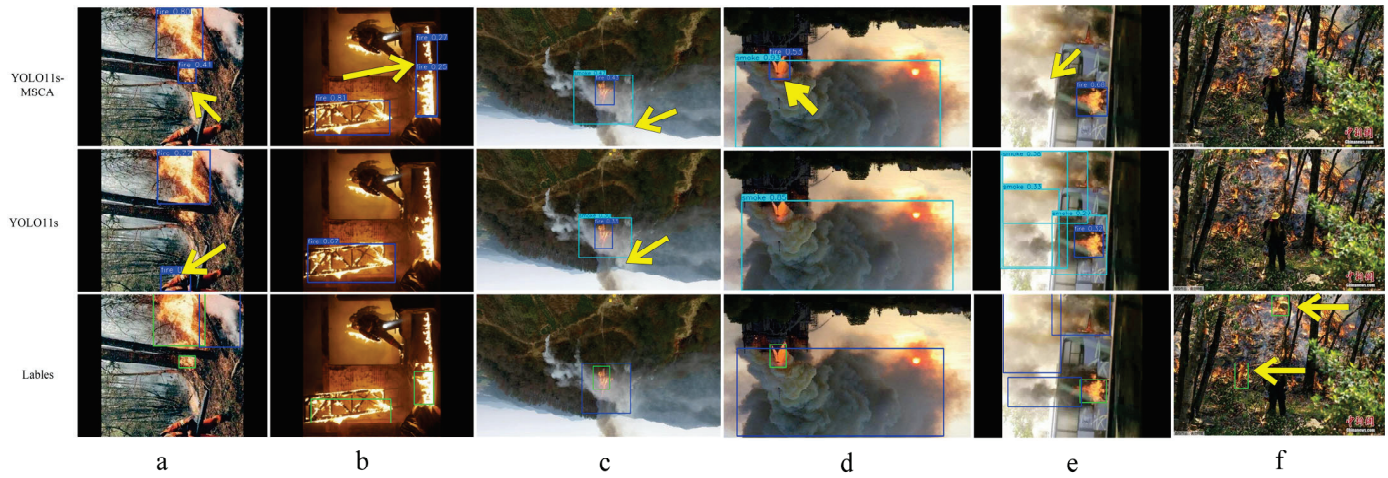


Figure 14. Analysis of typical examples of visual effects in YOLO11s-MSCA (CBM-Fire). (**a**). small target recognition. (**b**). precise localization of fire. (**c**). precise localization of smoke. (**d**). small-target fire. (**e**). large-area smoke dispersion. (**f**). occluded targets in complex environments.

Experimental results on D-Fire, Fire and Smoke Dataset, and CBM-Fire show that the improved YOLO11s-MSCA model has excellent and stable performance in flame and smoke target detection tasks. The model achieves high inference speed while maintaining high detection accuracy and achieves a good balance between accuracy and speed. Compared with mainstream models such as YOLOv9, YOLOv10, and RT-DETR, YOLO11s-MSCA has significant advantages in many indicators. By introducing Multi-Scale Convolutional Attention (MSCA) mechanism, the model has improved the ability of small target detection significantly and shows stronger robustness in complex environments. Although there are some limitations in extreme scenarios such as full smoke background or severe occlusion, the model shows consistent and excellent performance across multiple datasets, which verifies its good generalization ability and practical application potential.

Datasets produced by different teams differ, and these differences have an impact on model training and evaluation results. For example, the D-Fire dataset emphasizes its advantages such as large-scale, accurate labeling, diverse environment, and strong applicability and also points out its limitations such as uneven data distribution and single resolution [64].

First of all, when the amount of data is different, there will be significant differences between the CNN-based model and the Transformer model. As shown in Table 8, there are significant differences between the results of RT-DETR and YOLO series. When there was little difference in the amount of data, the Fire and Smoke Dataset with more resolutions and more scenes had an accurate recall rate of 3.4%, despite a 1.2% loss in accuracy. And the richer the scene, the more accurate the RT-DETR is, as shown in Tables 6 and 7. Meanwhile, the recall, mAP50, and mAP50-95 of the Fire and Smoke Dataset are significantly better than those of D-Fire, mainly because the Fire and Smoke Dataset directly divides interference items into one class instead of background as D-Fire had. The results show that the detection stability can be significantly improved by mixing the easily confused interference types (such as smoke, clouds, false fire points).

4. Discussion

The YOLO family has attracted much attention due to the optimal balance between speed and accuracy and has undergone continuous evolution, from YOLOv1 to the latest YOLO11 [42]. Each generation of enhancements has focused on minimizing latency while maximizing detection accuracy, and since YOLOv6, further advancing the utility of target

detection to make it more suitable for running on a resource-constrained device [58]. Considering the accuracy, speed, and practical deployment requirements of fire and smoke target detection, many scholars have improved the YOLO family [31,54,65]. In this paper, we use the classical models of the YOLO family for comparison to verify that the latest YOLO11 is still superior for smoke and fire detection tasks, that is, it guarantees an optimal balance of accuracy, speed, and utility of the model for fire and smoke target detection tasks. To meet specific requirements such as deployment environments, hardware constraints, and performance goals, YOLO11 models can be exported as configurations in ONNX or other formats. YOLO11 will be another great option for our fire and smoke target detection.

Considering the characteristics of fire and smoke target detection tasks, this research first considers the accuracy, while taking into account the speed (real-time) is not reduced and the model complexity (deployment practicality) is not increased to enhance the YOLO11 model. For the size variability of fire and smoke, the MSCA mechanism is introduced to pay attention to the multi-scale context information in the image. The loss plot and accuracy results demonstrate that the enhancement algorithm has improved the precision, recall, and confidence scores, while also achieving faster convergence during the training process. The overall indicators show that YOLO11s-MSCA improves the accuracy of the model without significantly increasing the complexity of the model. Compared with other YOLO series networks, the improved model network shows the best performance, and the accuracy of smoke objects is higher than that of fire objects. YOLO11s-MSCA has high accuracy and sensitivity for the detection of early smoke objects.

As observed by the visual effect map, the improved model can pay more attention to objects of different sizes in the image, especially those with small targets, has a better understanding of context, and can accurately distinguish some targets not recognized by the baseline model from the background. And the improvements and conclusions still hold in the Fire and Smoke and CBM-Fire datasets, where the model improvements possess generalizability and robustness to changes in complex environments. Nevertheless, the recognition of occluded targets and the understanding of complex backgrounds are still problems that have not fully been solved, and in the future, we can realize the lightweight deployment of the model by increasing the complexity of the dataset, improve the model by introducing other modules to enhance the model, or reduce the parameters of the model so that the model can provide fire and smoke detection services in practice, such as unmanned aerial vehicle (UAV) devices and camera devices, in order to reduce the time consumption and economic cost of data transmission.

At the same time, combining the conclusions of different datasets, it is concluded that the dataset construction strategy (data volume, scene diversity, classification system) will significantly affect the model performance, and reasonable classification design of interference items plays a key role in improving the detection stability. Clarifying the core objectives of model tasks and identifying tendencies and deployment scenarios are the prerequisites for formulating effective dataset construction strategies.

Author Contributions: Conceptualization, Y.L.(Yuxuan Li); methodology, Y.L. (Yuxuan Li) and L.N.; software, Y.L. (Yuxuan Li); validation, Y.L. (Yuxuan Li), H.F., F.Z. and L.W.; formal analysis, Y.L. (Yuxuan Li); investigation, Y.L. (Yuxuan Li); resources, Y.L. (Yuxuan Li); data curation, Y.L. (Yuxuan Li); writing—original draft preparation, Y.L. (Yuxuan Li); writing—review and editing, Y.L. (Yuxuan Li), H.F., F.Z., L.N., Y.L. (YunLiu), N.C., Q.D. and L.W.; visualization, Y.L. (Yuxuan Li); supervision, Y.L. (Yuxuan Li), F.Z. and L.W.; project administration, Q.D., F.Z. and L.W.; funding acquisition, Q.D., F.Z. and L.W.; All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China under Grant 32160369 and Grant 31860182, by Yunnan Provincial Key Science and Technology Program

under Grant 202202AD080010, and by the Ten Thousand Talent Plans for Young Top-Notch Talents of Yunnan Province under Grant YNWR-QNBJ-2019-026.

Institutional Review Board Statement: The study in this paper did not involve humans or animals.

Informed Consent Statement: Not applicable.

Data Availability Statement: The full implementation code, training model, and dataset download are publicly available at https://github.com/AbusAlex/YOLO11s-MSCA (accessed on 15 April 2025).

Conflicts of Interest: Author Fangrong Zhou was employed by the company Joint Laboratory of Power Remote Sensing Technology (Electric Power Research Institute, Yunnan Power Grid Company Ltd., China Southern Power Grid). 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. Morales-Hidalgo, D.; Oswalt, S.N.; Somanathan, E. Status and Trends in Global Primary Forest, Protected Areas, and Areas Designated for Conservation of Biodiversity from the Global Forest Resources Assessment 2015. *For. Ecol. Manag.* **2015**, *352*, 68–77. [CrossRef]
2. Sahoo, G.; Wani, A.; Rout, S.; Sharma, A.; Prusty, A.K. Impact and Contribution of Forest in Mitigating Global Climate Change. *Des. Eng.* **2021**, *4*, 667–682.
3. Lu, N. Dark Convolutional Neural Network for Forest Smoke Detection and Localization Based on Single Image. *Soft Comput.* **2022**, *26*, 8647–8659. [CrossRef]
4. Tien Bui, D.; Le, H.V.; Hoang, N.-D. GIS-Based Spatial Prediction of Tropical Forest Fire Danger Using a New Hybrid Machine Learning Method. *Ecol. Inform.* **2018**, *48*, 104–116. [CrossRef]
5. Pourmohamad, Y.; Abatzoglou, J.T.; Fleishman, E.; Short, K.C.; Shuman, J.; AghaKouchak, A.; Williamson, M.; Seydi, S.T.; Sadegh, M. Inference of Wildfire Causes From Their Physical, Biological, Social and Management Attributes. *Earth's Future* **2025**, *13*, e2024EF005187. [CrossRef]
6. Wah, W.; Gelaw, A.; Glass, D.C.; Sim, M.R.; Hoy, R.F.; Berecki-Gisolf, J.; Walker-Bone, K. Systematic Review of Impacts of Occupational Exposure to Wildfire Smoke on Respiratory Function, Symptoms, Measures and Diseases. *Int. J. Hyg. Environ. Health* **2025**, *263*, 114463. [CrossRef]
7. Zeng, B.; Zhou, Z.; Zhou, Y.; He, D.; Liao, Z.; Jin, Z.; Zhou, Y.; Yi, K.; Xie, Y.; Zhang, W. An Insulator Target Detection Algorithm Based on Improved YOLOv5. *Sci. Rep.* **2025**, *15*, 496. [CrossRef]
8. Zhang, F.; Zhao, P.; Xu, S.; Wu, Y.; Yang, X.; Zhang, Y. Integrating Multiple Factors to Optimize Watchtower Deployment for Wildfire Detection. *Sci. Total Environ.* **2020**, *737*, 139561. [CrossRef] [PubMed]
9. 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]
10. Shoshe, M.A.M.S.; Rahman, M.A. Improvement of Heat and Smoke Confinement Using Air Curtains in Informal Shopping Malls. *J. Build. Eng.* **2022**, *46*, 103676. [CrossRef]
11. Kuznetsov, G.V.; Volkov, R.S.; Sviridenko, A.S.; Strizhak, P.A. Fire Detection and Suppression in Rooms with Different Geometries. *J. Build. Eng.* **2024**, *90*, 109427. [CrossRef]
12. 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]
13. Khan, Z.A.; Ullah, F.U.M.; Yar, H.; Ullah, W.; Khan, N.; Kim, M.J.; Baik, S.W. Optimized Cross-Module Attention Network and Medium-Scale Dataset for Effective Fire Detection. *Pattern Recognit.* **2025**, *161*, 111273. [CrossRef]
14. Gaur, A.; Singh, A.; Kumar, A.; Kulkarni, K.S.; Lala, S.; Kapoor, K.; Srivastava, V.; Kumar, A.; Mukhopadhyay, S.C. Fire Sensing Technologies: A Review. *IEEE Sens. J.* **2019**, *19*, 3191–3202. [CrossRef]
15. Cheng, G.; Chen, X.; Wang, C.; Li, X.; Xian, B.; Yu, H. Visual Fire Detection Using Deep Learning: A Survey. *Neurocomputing* **2024**, *596*, 127975. [CrossRef]
16. 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]

17. 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]
18. Han, Z.; Tian, Y.; Zheng, C.; Zhao, F. Forest Fire Smoke Detection Based on Multiple Color Spaces Deep Feature Fusion. *Forests* **2024**, *15*, 689. [CrossRef]
19. Özel, B.; Alam, M.S.; Khan, M.U. Review of Modern Forest Fire Detection Techniques: Innovations in Image Processing and Deep Learning. *Information* **2024**, *15*, 538. [CrossRef]
20. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
21. Girshick, R. Fast R-CNN. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
22. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In Proceedings of the Advances in Neural Information Processing Systems; Curran Associates, Inc., Red Hook, NY, USA; 2015; Volume 28.
23. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In *Computer Vision—ECCV 2016, Proceedings of the 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016*; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Springer International Publishing: Cham, Switzerland, 2016; pp. 21–37.
24. Ragab, M.G.; Abdulkadir, S.J.; Muneer, A.; Alqushaibi, A.; Sumiea, E.H.; Qureshi, R.; Al-Selwi, S.M.; Alhussian, H. A Comprehensive Systematic Review of YOLO for Medical Object Detection (2018 to 2023). *IEEE Access* **2024**, *12*, 57815–57836. [CrossRef]
25. Zhao, Z.-Q.; Zheng, P.; Xu, S.-T.; Wu, X. Object Detection With Deep Learning: A Review. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *30*, 3212–3232. [CrossRef] [PubMed]
26. Mamadaliev, D.; Touko, P.L.M.; Kim, J.-H.; Kim, S.-C. ESFD-YOLOv8n: Early Smoke and Fire Detection Method Based on an Improved YOLOv8n Model. *Fire* **2024**, *7*, 303. [CrossRef]
27. Wang, D.; Qian, Y.; Lu, J.; Wang, P.; Yang, D.; Yan, T. Ea-Yolo: Efficient Extraction and Aggregation Mechanism of YOLO for Fire Detection. *Multimed. Syst.* **2024**, *30*, 287. [CrossRef]
28. Jiao, Z.; Zhang, Y.; Xin, J.; Mu, L.; Yi, Y.; Liu, H.; Liu, D. A Deep Learning Based Forest Fire Detection Approach Using UAV and YOLOv3. In Proceedings of the 2019 1st International Conference on Industrial Artificial Intelligence (IAI), Shenyang, China, 23–27 July 2019; pp. 1–5.
29. Wu, H.; Hu, Y.; Wang, W.; Mei, X.; Xian, J. Ship Fire Detection Based on an Improved YOLO Algorithm with a Lightweight Convolutional Neural Network Model. *Sensors* **2022**, *22*, 7420. [CrossRef]
30. Xue, Q.; Lin, H.; Wang, F. FCDM: An Improved Forest Fire Classification and Detection Model Based on YOLOv5. *Forests* **2022**, *13*, 2129. [CrossRef]
31. Chen, X.; Xue, Y.; Hou, Q.; Fu, Y.; Zhu, Y. RepVGG-YOLOv7: A Modified YOLOv7 for Fire Smoke Detection. *Fire* **2023**, *6*, 383. [CrossRef]
32. Zhang, T.; Wang, F.; Wang, W.; Zhao, Q.; Ning, W.; Wu, H. Research on Fire Smoke Detection Algorithm Based on Improved YOLOv8. *IEEE Access* **2024**, *12*, 117354–117362. [CrossRef]
33. Zhao, C.; Zhao, L.; Zhang, K.; Ren, Y.; Chen, H.; Sheng, Y. Smoke and Fire-You Only Look Once: A Lightweight Deep Learning Model for Video Smoke and Flame Detection in Natural Scenes. *Fire* **2025**, *8*, 104. [CrossRef]
34. Liu, H.; Zhang, F.; Xu, Y.; Wang, J.; Lu, H.; Wei, W.; Zhu, J. TFNet: Transformer-Based Multi-Scale Feature Fusion Forest Fire Image Detection Network. *Fire* **2025**, *8*, 59. [CrossRef]
35. Safarov, F.; Muksimova, S.; Kamoliddin, M.; Cho, Y.I. Fire and Smoke Detection in Complex Environments. *Fire* **2024**, *7*, 389. [CrossRef]
36. Alkhammash, E.H. Multi-Classification Using YOLOv11 and Hybrid YOLO11n-MobileNet Models: A Fire Classes Case Study. *Fire* **2025**, *8*, 17. [CrossRef]
37. de Venâncio, P.V.A.B.; Lisboa, A.C.; Barbosa, A.V. An Automatic Fire Detection System Based on Deep Convolutional Neural Networks for Low-Power, Resource-Constrained Devices. *Neural Comput. Applic* **2022**, *34*, 15349–15368. [CrossRef]
38. Catargiu, C.; Cleju, N.; Ciocoiu, I.B. A Comparative Performance Evaluation of YOLO-Type Detectors on a New Open Fire and Smoke Dataset. *Sensors* **2024**, *24*, 5597. [CrossRef] [PubMed]
39. Geng, X.; Han, X.; Cao, X.; Su, Y.; Shu, D. YOLOV9-CBM: An Improved Fire Detection Algorithm Based on YOLOV9. *IEEE Access* **2025**, *13*, 19612–19623. [CrossRef]
40. Terven, J.; Córdova-Esparza, D.-M.; Romero-González, J.-A. A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS. *Mach. Learn. Knowl. Extr.* **2023**, *5*, 1680–1716. [CrossRef]

41. Lin, T.-Y.; Dollar, 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. 2117–2125.
42. Ali, M.L.; Zhang, Z. The YOLO Framework: A Comprehensive Review of Evolution, Applications, and Benchmarks in Object Detection. *Computers* **2024**, *13*, 336. [CrossRef]
43. Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In Proceedings of the Proceedings of the 32nd International Conference on Machine Learning, PMLR, Lille, France, 6–11 July 2015; pp. 448–456.
44. Elfwing, S.; Uchibe, E.; Doya, K. Sigmoid-Weighted Linear Units for Neural Network Function Approximation in Reinforcement Learning. *Neural Netw.* **2018**, *107*, 3–11. [CrossRef]
45. Liu, W.; Liao, S.; Ren, W.; Hu, W.; Yu, Y. High-Level Semantic Feature Detection: A New Perspective for Pedestrian Detection. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 5187–5196.
46. Zhang, H.; Zu, K.; Lu, J.; Zou, Y.; Meng, D. EPSANet: An Efficient Pyramid Split Attention Block on Convolutional Neural Network. *arXiv* **2021**, arXiv:2105.14447.
47. Wang, K.; Liew, J.H.; Zou, Y.; Zhou, D.; Feng, J. PANet: Few-Shot Image Semantic Segmentation With Prototype Alignment. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9197–9206.
48. Zhang, D. A Yolo-based Approach for Fire and Smoke Detection in IoT Surveillance Systems | EBSCOhost. Available online: https://openurl.ebsco.com/contentitem/doi:10.14569/ijacsa.2024.0150109?sid=ebsco:plink:crawler&id=ebsco:doi:10.14569/ijacsa.2024.0150109 (accessed on 6 March 2025).
49. Guo, M.-H.; Lu, C.-Z.; Hou, Q.; Liu, Z.; Cheng, M.-M.; Hu, S. SegNeXt: Rethinking Convolutional Attention Design for Semantic Segmentation. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 1140–1156.
50. Chen, L.; Zhang, H.; Xiao, J.; Nie, L.; Shao, J.; Liu, W.; Chua, T.-S. SCA-CNN: Spatial and Channel-Wise Attention in Convolutional Networks for Image Captioning. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 5659–5667.
51. Han, Q.; Fan, Z.; Dai, Q.; Sun, L.; Cheng, M.-M.; Liu, J.; Wang, J. On the Connection between Local Attention and Dynamic Depth-Wise Convolution. *arXiv* **2022**, arXiv:10.48550/arXiv.2106.04263.
52. Peng, C.; Zhang, X.; Yu, G.; Luo, G.; Sun, J. Large Kernel Matters—Improve Semantic Segmentation by Global Convolutional Network. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 4353–4361.
53. Lin, M.; Chen, Q.; Yan, S. Network In Network 2014. *arXiv* **2022**, arXiv:10.48550/arXiv.1312.4400.
54. Pan, W.; Wang, X.; Huan, W. EFA-YOLO: An Efficient Feature Attention Model for Fire and Flame Detection. *arXiv* **2024**, arXiv:2409.12635.
55. Jiang, K.; Xie, T.; Yan, R.; Wen, X.; Li, D.; Jiang, H.; Jiang, N.; Feng, L.; Duan, X.; Wang, J. An Attention Mechanism-Improved YOLOv7 Object Detection Algorithm for Hemp Duck Count Estimation. *Agriculture* **2022**, *12*, 1659. [CrossRef]
56. Cao, L.; Shen, Z.; Xu, S. Efficient Forest Fire Detection Based on an Improved YOLO Model. *Vis. Intell.* **2024**, *2*, 20. [CrossRef]
57. Jia, X.; Tong, Y.; Qiao, H.; Li, M.; Tong, J.; Liang, B. Fast and Accurate Object Detector for Autonomous Driving Based on Improved YOLOv5. *Sci. Rep.* **2023**, *13*, 9711. [CrossRef]
58. Li, C.; Li, L.; Jiang, H.; Weng, K.; Geng, Y.; Li, L.; Ke, Z.; Li, Q.; Cheng, M.; Nie, W.; et al. YOLOv6: A Single-Stage Object Detection Framework for Industrial Applications. *arXiv* **2022**, arXiv:2209.02976.
59. Wang, C.-Y.; Bochkovskiy, A.; Liao, H.-Y.M. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475.
60. Al Mudawi, N.; Qureshi, A.M.; Abdelhaq, M.; Alshahrani, A.; Alazeb, A.; Alonazi, M.; Algarni, A. Vehicle Detection and Classification via YOLOv8 and Deep Belief Network over Aerial Image Sequences. *Sustainability* **2023**, *15*, 14597. [CrossRef]
61. Wang, C.-Y.; Yeh, I.-H.; Mark Liao, H.-Y. YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information. In *Computer Vision—ECCV 2024, Proceedings of the 18th European Conference, Milan, Italy, 29 September–4 October 2024*; Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G., Eds.; Springer Nature Switzerland: Cham, Switzerland, 2025; pp. 1–21.
62. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; Ding, G. YOLOv10: Real-Time End-to-End Object Detection. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 107984–108011.
63. Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; Chen, J. DETRs Beat YOLOs on Real-Time Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 16965–16974.

64. Boroujeni, S.P.H.; Mehrabi, N.; Afghah, F.; McGrath, C.P.; Bhatkar, D.; Biradar, M.A.; Razi, A. Fire and Smoke Datasets in 20 Years: An In-Depth Review. *arXiv* **2025**, arXiv:2503.14552.
65. Saydirasulovich, S.N.; Mukhiddinov, M.; Djuraev, O.; Abdusalomov, A.; Cho, Y.-I. An Improved Wildfire Smoke Detection Based on YOLOv8 and UAV Images. *Sensors* **2023**, *23*, 8374. [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

Geographically Weighted Random Forest Based on Spatial Factor Optimization for the Assessment of Landslide Susceptibility

Feifan Lu [1,†], Guifang Zhang [1,2,3,*,†], Tonghao Wang [1], Yumeng Ye [1] and Qinghao Zhao [1]

[1] School of Earth Sciences and Engineering, Sun Yat-Sen University, Zhuhai 519082, China; luff@mail2.sysu.edu.cn (F.L.); wangth6@mail2.sysu.edu.cn (T.W.); yeym9@mail2.sysu.edu.cn (Y.Y.); zhaoqh8@mail2.sysu.edu.cn (Q.Z.)
[2] Guangdong Provincial Key Laboratory of Geological Processes and Mineral Resources, Zhuhai 519082, China
[3] Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai), Zhuhai 519082, China
* Correspondence: zhgfang@mail.sysu.edu.cn
† These authors contributed equally to this work.

Abstract: Landslide susceptibility mapping is a crucial tool for landslide disaster risk management. However, the spatial heterogeneity of landslide conditioning factors affects the accuracy of predictions. This study proposes a novel method combining GeoDetector and geographical weighted random forest (GeoD-GWRF), a local machine learning approach. The GeoD-GWRF model can select landslide conditioning factors from the perspective of spatial differentiation and interpret the influence of factors on landslides at a local scale. The model's applicability is verified using Luhe County, Guangdong Province, as a case study. Compared to the traditional random forest model, the GeoD-GWRF model achieves higher prediction accuracy (AUC = 0.942). In addition, the model is applicable to broader study areas and can provide more targeted prediction results. This method offers a valuable reference for exploring spatial heterogeneity in landslide susceptibility mapping.

Keywords: landslide susceptibility mapping; GeoDetector; geographical weighted random forest; spatial heterogeneity

1. Introduction

Landslides are frequent geological disasters worldwide, significantly impacting the natural environment, human society, and economic development [1,2]. Research indicates that from 2004 to 2016, there were 4862 landslide events globally, resulting in 55,997 fatalities, with Asia being the most affected region [3]. Landslide susceptibility mapping (LSM), as a crucial tool for enhancing landslide risk management and mitigating disaster losses, delineates the spatial probability of landslide occurrences based on local topographic conditions [4,5]. However, the spatial heterogeneity of landslide conditioning factors (LCFs) within the study area may increase the uncertainty of LSM results, making it essential to enhance the accuracy of landslide susceptibility assessments.

Large-scale LSM methods can generally be classified into qualitative and quantitative categories. Qualitative analysis, such as the analytic hierarchy process (AHP) [6,7], assigns subjective weights to LCFs based on expert knowledge. This approach lacks the ability to provide objective quantitative evaluations, often carrying inherent subjectivity and being susceptible to human interference [8]. In quantitative methods, statistical models such as information value (IV) [9], frequency ratio (FR) [10], and logistic regression (LR) [11] are widely used due to their interpretability and ease of implementation. However, these

models often assume that the relationship between LCFs and landslides is either normal or linear, which limits their accuracy. In contrast, machine learning (ML) models demonstrate greater applicability in LSM due to their powerful capabilities in handling nonlinear and complex data [12,13].

Therefore, ML models, such as support vector machines (SVM) [14], random forests (RF) [15], and artificial neural networks (ANN) [16], dominate LSM. Further research indicates that RF has significant advantages in handling high-dimensional data and preventing overfitting [17,18]. This tree-based ensemble algorithm achieves higher robustness and accuracy in LSM with minimal adjustments required before model training [19]. Although ML models, represented by RF, have achieved remarkable success in LSM, the training of these models typically assumes that the samples are independent and identically distributed. This leads to the assumption that the influence of each LCF on landslides is spatially constant. Global models overlook the spatial variability of LCFs across different geographic locations within the study area, thus failing to account for the spatial non-stationarity between landslides and factors [20,21]. The consideration of spatial heterogeneity in LCFs has led to the gradual integration of local regression concepts into LSM [22,23]. Furthermore, Zhao et al. [24] combined geographically weighted methods with neural networks, whose results showed that this approach may effectively address the issue of local overfitting observed in traditional ML models.

In addition to insufficient consideration of spatial structure, the lack of effective methods for selecting LCFs is also a major cause of reduced accuracy in LSM [25]. LCFs can be broadly divided into two categories: static or environmental factors (e.g., topographic conditions, geological conditions, and land cover) and dynamic or triggering factors (e.g., precipitation and earthquakes) [26,27]. Reichenbach et al. [28] conducted a statistical analysis of LSM studies and identified a total of 596 factors. However, there are no universally accepted guidelines regarding the number of factors to use or the methods for selecting LCFs. Redundant and noisy conditioning factors can increase model uncertainty and reduce prediction accuracy. Moreover, overfitting in ML models may be a result of excessive data [29]. Common factor selection methods, such as Pearson correlation coefficient [30], principal component analysis [31], collinearity tests [32], and recursive elimination [18], have optimized the data volume to some extent and improved the reliability of LSM. However, these methods overlook the spatial information of landslides and LCFs, making it impossible to capture the spatial distribution patterns of LCFs and their relationship to landslides. GeoDetector, with its exceptional ability to explore spatial heterogeneity of factors [33], takes into account how the spatial variation of landslides can lead to different responses in the dependent variables [34,35].

This study aims to address the limitations in research on spatial structure in LSM by adopting a hybrid method using GeoDetector and the geographically weighted random forest (GWRF) model (GeoD-GWRF). By fully considering the spatial heterogeneity of factors, the most suitable LCFs are selected. A local machine learning approach that incorporates spatial structure is then used for landslide susceptibility assessment. The goal is to provide new methods and perspectives for understanding the interaction patterns between LCFs and landslides, as well as for disaster prevention and mitigation.

2. Study Area and Materials

2.1. Study Area

The study area, Luhe County, covers 986 km^2 on the southeastern coast of Guangdong Province, located between 115°24′ to 115°49′ E and 23°08′ to 23°28′ N (Figure 1). The topography is characterized by higher elevations in the east and west, with a lower central region. The central and southern areas consist of river terraces and alluvial plains, while

the northwestern and southeastern areas are dominated by medium-low mountains. The elevation in the study area ranges from 12 to 1219 m, with mountainous areas above 500 m covering 263.6 km^2. Luhe County is situated in a subtropical monsoon climate zone, with abundant rainfall. The average annual precipitation ranges from 1800 to 2400 mm, with the maximum annual rainfall reaching up to 3728 mm. The lithology of the study area can be classified into five types: single-layer soil, double-layer soil, multi-layer soil, layered clastic rock, and massive intrusive rock. The Lianhua Mountain fault runs through the central region, with gentle slopes and a large scale. The river system is mainly formed by tributaries of the Luo and Rong rivers, with significant elevation drops in the riverbeds, providing abundant hydropower resources. Combined with active human engineering activities, these features lead to frequent landslides, posing a serious threat to public safety.

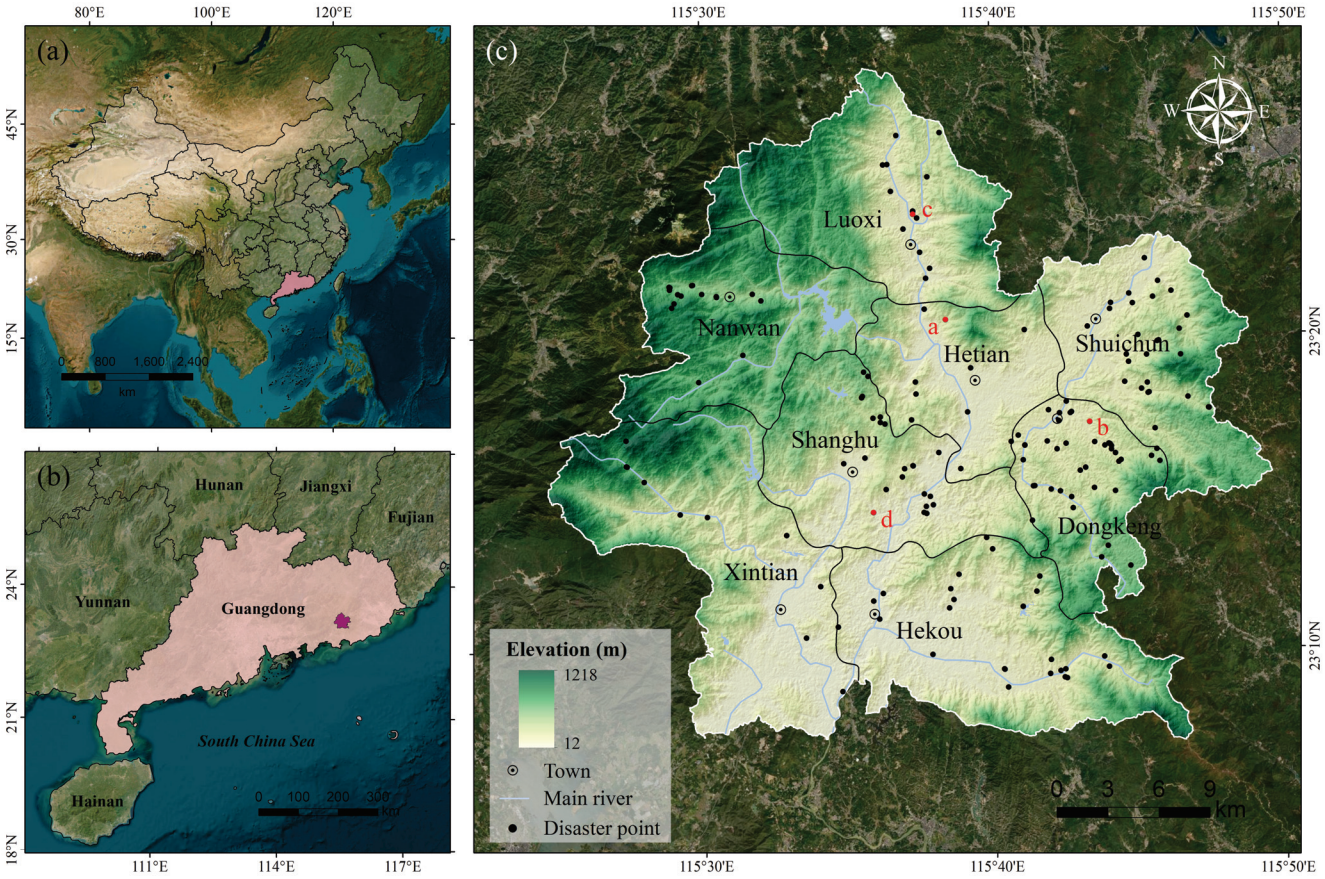


Figure 1. Basic information of the study area. (**a**) location in China; (**b**) location in Guangdong Province; (**c**) location and landslide distribution in Luhe County.

2.2. Data Source

Digital Elevation Model (DEM) data were used to process landslide driving factors, including slope. The land use data, covering ten major types—such as arable land, woodland, grassland, shrubland, wetland, water bodies, tundra, artificial surfaces, bare land, and glaciers—were sourced from GlobeLand30 (http://loess.geodata.cn). Upon comparison, the 2020 land use data for Luhe was found to closely resemble that of 2000, thus addressing concerns regarding data timeliness in relation to the landslide occurrence period [36]. Therefore, the 2020 land use data was chosen for its higher quality and closer alignment with the landslide survey timeline. Additionally, vector data for roads, rivers, and faults were used for distance-based analysis. A summary of the dataset characteristics is provided in Table 1.

Table 1. Key spatial data information required for this study.

Data Name	Data Format	Scale or Resolution	Data Source
DEM	GeoTIFF	12.5 m	ALOS GDEM
Disaster points	.shp (point)	1:50,000	Field work and remote sensing interpretation
Remote sensing image	GeoTIFF	30 m	L1T product of LANDSAT-8
Road	.shp (line)	1:50,000	Calibration data based on Amap
Water body	.shp (polygon)	1:50,000	Calibration data based on Amap
Fault	.shp (line)	1:50,000	Zijin regional geological map
Land use	GeoTIFF	30 m	GlobeLand30

The landslide inventory used in this study was primarily obtained from historical landslide records registered by the Luhe County Bureau of Natural Resources. These landslides were documented through field measurements and on-site investigations conducted by local government authorities. The inventory includes detailed information such as landslide names, affected areas, hazard scales, deformation and failure characteristics, disaster features, and their developmental trends. Additionally, the most recent geological hazard risk survey in Luhe County was leveraged to further supplement and verify the landslide data. This effort involved 1:10,000 remote sensing interpretation across a 986 km^2 area, with disaster and hazard points systematically reviewed. Field engineering investigations were also conducted using administrative villages as the basic spatial unit, resulting in over 80% of the landslide points being verified on-site. Ultimately, a total of 166 landslides were confirmed. Among these, 124 were classified as small-scale landslides ($<10 \times 10^4$ m^3), while 42 were medium-scale (10×10^4–100×10^4 m^3). The majority of the identified landslides are soil-type, with gravel landslides being less common. Based on lithological characteristics, landslides were categorized into accumulation landslides and clay landslides. Accumulation landslides mainly consist of quaternary deposits or anthropogenic accumulations such as gravel and slag, while clay landslides are primarily composed of clayey soils with gravels, silty clay mixed with pebbles, and fragmented rock. In terms of morphology, the landslides generally exhibit tongue-shaped planar features and concave longitudinal profiles. Spatially, they are predominantly concentrated in the low mountainous and hilly areas of Nanwan, Shanghu, Dongkeng, and Shuichun towns. Figure 2 illustrates examples of four representative landslides in Luhe County, as marked in Figure 1c. These images were captured by UAV during field verification.

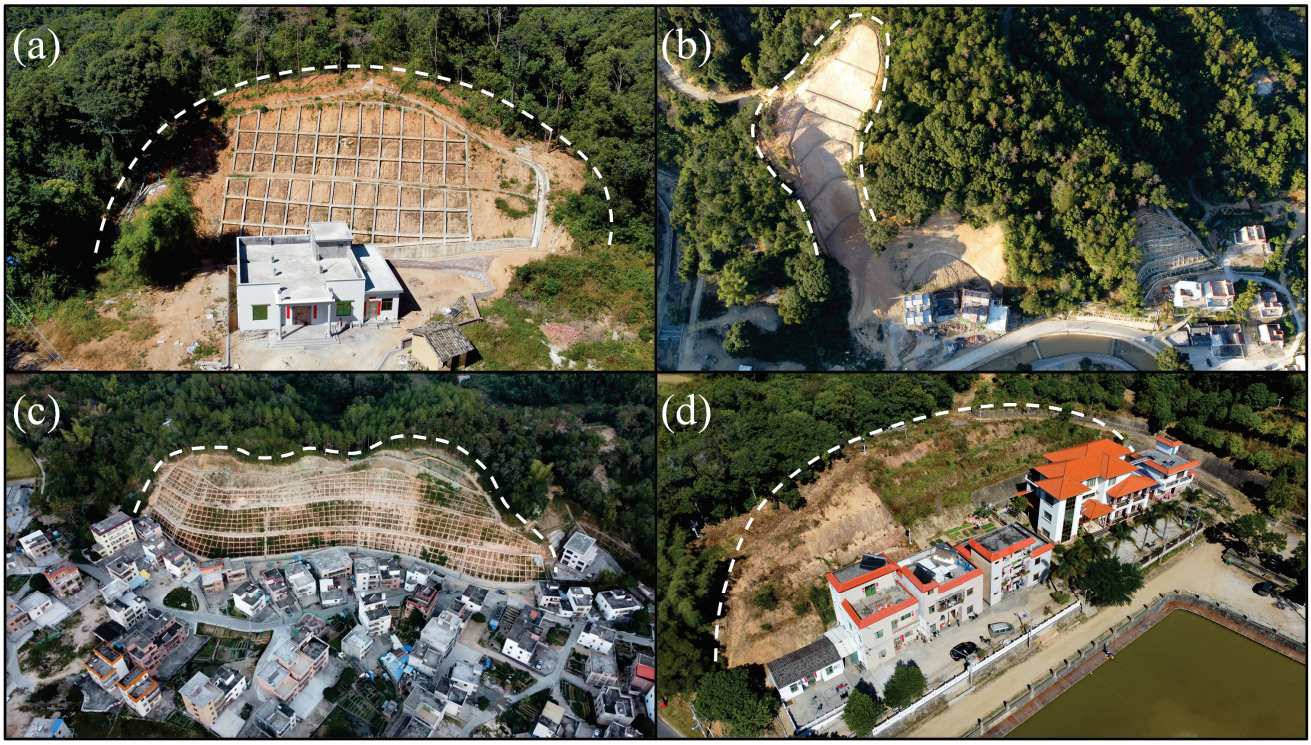


Figure 2. Landslide examples from field investigation and their locations marked in Figure 1c. (**a**) a small-scale landslide in Hetian; (**b**) a landslide adjacent to a road in Dongkeng; (**c**) a medium-sized landslide in Luoxi; (**d**) a landslide close to buildings in Shanghu.

2.3. Landslide Conditioning Factors

Landslide occurrence is influenced by both natural and human factors. Precipitation, particularly intense short-term rainfall is often considered a triggering factor due to its immediate impact on landslide stability. Such studies typically involve collecting accurate precipitation amounts before and after landslide events, as well as determining the number of landslides triggered by precipitation [37,38]. In this study, we utilized the average annual precipitation data from 2003 to 2022 (with a resolution of 1000 m). The average annual precipitation accounts for the lag effect of precipitation on landslides, as it may not immediately trigger a landslide. Instead, landslides are more likely to occur after prolonged water infiltration, when the soil saturation reaches a critical level. As a result, average annual precipitation was included as one of the environmental factors to better capture its spatial variability in influencing landslide occurrences, in conjunction with other factors. Ultimately, considering the geological and environmental characteristics of the study area, 15 LCFs were preliminarily selected, including geological, topographical, hydrological, and human activity factors: elevation, slope, curvature, terrain relief (TR), surface roughness (SR), normalized difference vegetation index (NDVI), precipitation, slope of aspect (SOA), terrain wetness index (TWI), stream power index (SPI), distance to fault (DTF), distance to road (DTR), distance to water (DTW), lithology, and land use.

Considering the study area coverage and the computational complexity of the model, all raster layers were standardized to a spatial resolution of 30 m × 30 m. Data processing and visualization were performed using ERSI ArcGIS 10.8 and ENVI 5.6 software. The thematic layers of various LCFs are shown in Figure 3.

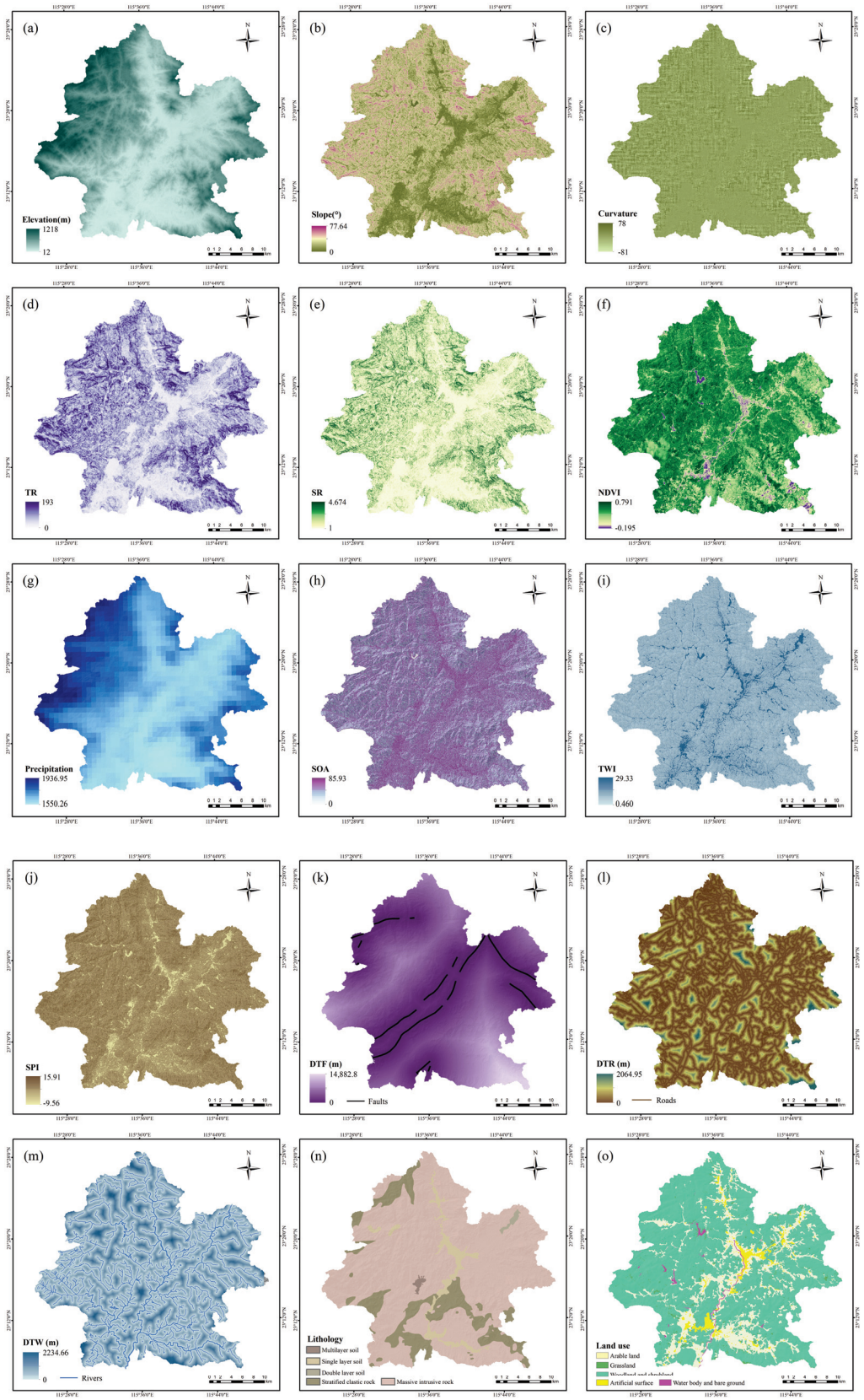


Figure 3. Thematic map layers of landslide conditioning factors. (**a**) elevation; (**b**) slope; (**c**) curvature; (**d**) terrain relief (TR); (**e**) surface roughness (SR); (**f**) normalized difference vegetation index (NDVI); (**g**) precipitation; (**h**) slope of aspect (SOA); (**i**) terrain wetness index (TWI); (**j**) stream power index (SPI); (**k**) distance to fault (DTF); (**l**) distance to road (DTR); (**m**) distance to water (DTW); (**n**) lithology; (**o**) land use.

3. Methodology

Figure 4 illustrates the main workflow of this study: (1) collecting the landslide inventory of Luhe County and initially selecting 15 landslide conditioning factors to construct a GIS dataset; (2) applying GeoDetector to perform factor selection based on optimal reclassification; (3) inputting the selected factors into RF and GWRF models to construct the GeoD-RF and GeoD-GWRF models; (4) comparing and evaluating the performance of these hybrid optimized models; and (5) exploring the spatial relationships between LCFs and landslides.

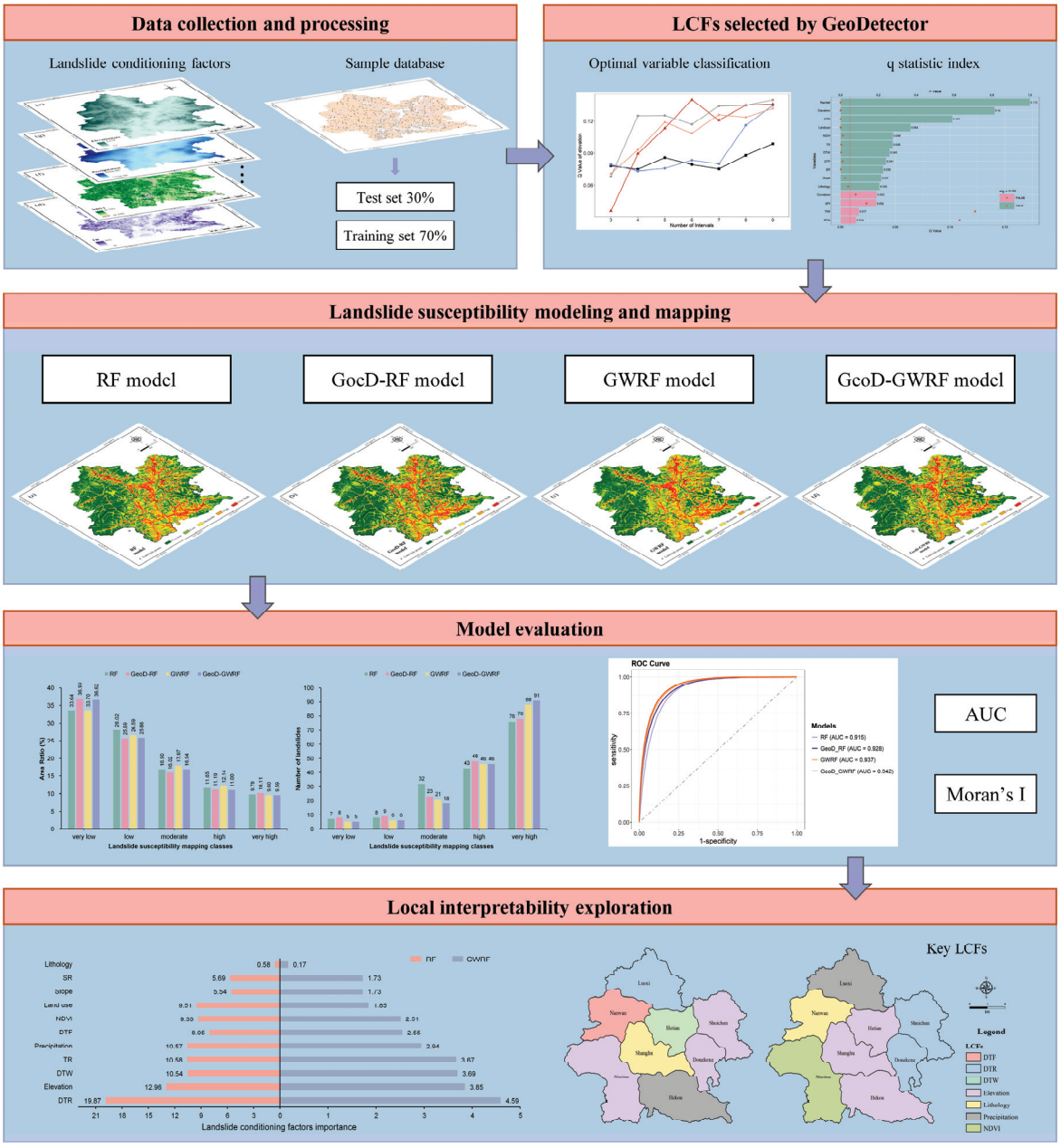


Figure 4. The main workflow framework of this study.

3.1. Preparation of the Sample Set

For large-scale study areas with limited landslide inventories, it is common to represent each landslide by its centroid, a method shown to be effective in previous work [39,40]. In supervised machine-learning models, sample preparation is indispensable: non-landslide (negative) samples are as critical as landslide (positive) samples in regional susceptibility

mapping. Previous studies typically define buffer zones at a fixed distance around known landslide sites and then randomly select negative samples from outside these buffers [41,42]. The selection of non-landslide samples by this method is highly subjective and random, reflecting only that no landslides have occurred at these locations to date; however, because potential landslides continually evolve, these points may still experience failure in the future [43]. Other studies have opted to draw negative samples from specific areas (such as zones of low slope), but this strategy concentrates negatives in geologically homogeneous regions, thereby overlooking other influential factors associated with landslides and potentially biasing model performance [44]. To improve the quality of negative–sample selection, Hu et al. (2020) [45] developed a fractal-theory–based method targeting very-low susceptibility areas, which outperformed traditional sampling from low-slope and landslide-free zones. Khabiri et al. (2023) [44] introduced the PISA-m classification framework, integrating geotechnical parameters into the sampling process and cautioning that conventional non-landslide–zone sampling can induce model overfitting. In their recent work, Dou et al. (2023) [46] applied an information-value model to partition susceptibility zones and demonstrated that drawing negative samples from low and very-low susceptibility regions substantially enhances model performance.

The various methods for selecting non-landslide samples described above have been demonstrated effective. Considering the study area's characteristics as well as the completeness and quality of collected data, we decided to employ the information value model to determine landslide negative samples. Based on probability theory, the information value model is a commonly used statistical method in landslide susceptibility mapping, assigning a weight to each categorical unit and using the resulting entropy value to represent the probability of geological disasters; the larger the information value, the greater the likelihood of landslide occurrence within that unit. By characterizing factor influence through entropy magnitude, this approach offers a simple and effective means of representing the combined contribution of multiple factors to landslide susceptibility. Selecting negative samples from regions with low information values improves the precision of chosen samples and avoids dominance by any single influencing factor. The fundamental formula of the information value model is as follows:

$$IV_i = \ln\left(\frac{N_i/N}{S_i/S}\right) \tag{1}$$

$$IV = \sum_{i=1}^{n} IV_i = \sum_{i=1}^{n} \ln\left(\frac{N_i/N}{S_i/S}\right) \tag{2}$$

where S represents the total area of the study region, and N denotes the total number of landslides. S_i and N_i refer to the area and the number of landslides within the i-th interval of a given factor, respectively. IV_i is the information value of the i-th interval, while IV represents the total information value for the entire study area.

3.2. GeoDetector

GeoDetector is a spatial-based multivariate statistical model used for quantitatively evaluating the impact of a potential factor on a spatial phenomenon [47]. It makes minimal assumptions about the data, potentially overcoming the limitations of other statistical methods when dealing with spatial variables and providing clear physical interpretations [48]. The core assumption of GeoDetector in landslide studies is that if LCFs control or contribute to landslide occurrences, the spatial distribution of landslides should resemble the spatial distribution of these conditioning factors. Therefore, GeoDetector can reveal the underlying driving forces of a phenomenon by quantifying the influence of individual factors, making it suitable for factor selection. The factor detector within GeoDetector is used to assess the spatial heterogeneity of landslides (y) and evaluate the explanatory power of a LCF (x) on

this spatial heterogeneity. By performing overlay analysis on the y-layer and x-layer, local and global variances are calculated to detect whether the x-layer contributes to the spatial distribution of landslides (Figure 5b). If x and y are correlated, a similar spatial pattern will exist between the two (Figure 5c). The explanatory power is ultimately measured by the q-value, calculated as follows:

$$q = 1 - \frac{1}{N\sigma^2}\sum_{n=1}^{Z} N_n\sigma_n^2 = 1 - \frac{WSS}{TSS} \tag{3}$$

$$WSS = \sum_{n=1}^{Z} N_n\sigma_n^2 \tag{4}$$

$$TSS = N\sigma^2 \tag{5}$$

where Z represents the number of zones in the x-layer, while N_n and N denote the sample sizes in zone n and the entire study area, respectively. σ_n^2 and σ^2 correspond to the variance of the dependent variable in zone n and the entire study area. WSS is the sum of within-zone variances, and TSS is the total variance of the entire area. The range of the q is [0, 1], where a higher q value indicates a stronger explanatory power of the independent variable X on the spatial heterogeneity of the dependent variable Y, and vice versa. The GeoDetector analysis was implemented using the *GD* package (version 10.3) in R.

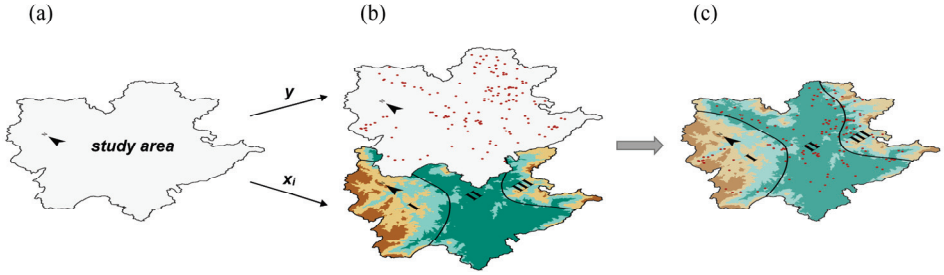


Figure 5. Principle of GeoDetector. By overlaying the y layer and x layer, the spatial similarity between independent variables and the dependent variable is obtained. (**a**) schematic of the study area; (**b**) overlay analysis of the dependent variable and the discretized independent variables; (**c**) spatial pattern analysis between the dependent variable and independent variables.

3.3. Random Forest Model

The random forest (RF) is an ensemble machine learning method proposed by Breiman (2001) [49]. It combines multiple Decision Trees (DTs) through a random sampling approach for classification and regression tasks. The working principles of RF, as illustrated in Figure 6c, can be summarized as follows: (1) A bootstrap sampling is performed on the original training dataset to randomly select n new sample subsets with replacement; (2) select a set of features randomly at each re-sampling, and build n decision trees on this basis; (3) The generated trees are then aggregated to form a random forest, where the final output is determined by combining the predictions of each DT. Each sample subset used to create a DT represents two-thirds of the total samples, while the remaining samples are referred to as the out-of-bag (OOB) set. By calculating the OOB error, the relative importance of the LCFs in the prediction results can be quantified. The formula for calculating LCF importance is as follows:

$$LCFi_x = \frac{1}{n}\sum_t \left(OOB^t_{MSE,\ perm_x} - OOB^t_{MSE}\right) \tag{6}$$

where $LCFi_x$ represents the importance of the landslide conditioning factor x, n is the total number of trees in the forest, OOB^t_{MSE} is the mean square error of tree t before x is permuted, and $OOB^t_{MSE,\ perm_x}$ is the mean square error of tree t after x is permuted.



Figure 6. The principle of Geographical Weighted Random Forest. By using bandwidth to decompose the global machine learning model, it identifies the optimal local RF model for prediction in each grid. (**a**) schematic of the bandwidth used in GWRF; (**b**) different RF sub-models applied in various regions; (**c**) working principle of each RF sub-model.

RF mitigates the tendency of a single DT to overfit and exhibits high tolerance to outliers and noise [50]. Due to its robustness and accuracy in handling complex data, the RF model has been widely applied in LSM, showing promising performance [12,17]. The RF model of this study was implemented using the *randomForest* package (version 4.7-1.1) in R.

3.4. Geographical Weighted Random Forest Model

Although RF shows high applicability in LSM, it remains a global model. The global model is based on the assumption that the relationship between factors and landslides is constant across the entire spatial domain, thereby overlooking the spatial non-stationarity of the variables. Geographical Weighted Random Forest (GWRF) model extends RF by incorporating spatial dependencies, allowing for the consideration of the spatial heterogeneity of the variables themselves [51]. The working principles of GWRF are similar to geographically weighted regression. Specifically, the GWRF model incorporates the spatial location information of variables into the RF model. By fitting variable datasets located at different spatial positions, it generates local RF sub-models for each spatial unit (Figure 6). Thus, the simplified equation for the GWRF model can be expressed as:

$$Y_i(\mu_i, \nu_i) = f(x_i, (\mu_i, \nu_i)) + \varepsilon_i \tag{7}$$

where $f(x_i, (\mu_i, \nu_i))$ represents the prediction of the local RF calibrated based on position i, and (μ_i, ν_i) is the coordinate of the spatial unit at location i.

Each local RF sub-model covers a spatial range of data known as its neighborhood, and the parameter controlling the size of the neighborhood is called bandwidth (Figure 6a). There are two types of bandwidth: adaptive and fixed. The adaptive bandwidth ensures that the number of samples in each neighborhood is consistent, while the fixed bandwidth generates neighborhoods of uniform spatial range. In this study, the adaptive bandwidth GWRF model was adopted because it allows the neighborhood size to vary, making it more flexible for landslide data with different spatial sampling densities [52]. The GWRF model of this study was implemented using the *SpatialML* package (version 0.1.7) in R.

3.5. Validation of Models

For landslides, a typical binary classification problem, the receiver operating characteristic (ROC) curve is an effective tool to evaluate the predictive ability of a model [53]. The ROC curve, derived from the confusion matrix, reflects the dynamic changes in classification results. The accuracy of the ROC curve is interpreted by calculating the area under the curve (AUC). The AUC ranges from [0.5, 1], with higher AUC values indicating better model performance [54]. The relevant formulas derived from the confusion matrix are as follows:

$$Sensitivity = \frac{TP}{TP + FN} \tag{8}$$

$$Specificity = \frac{TN}{FP + TN} \tag{9}$$

$$AUC = \frac{\sum TP + \sum TN}{P + N} \tag{10}$$

where TP (true positive) represents the number of correctly classified landslide samples, TN (true negative) represents the number of correctly classified non-landslide samples, FP (false positive) represents the number of landslide samples incorrectly classified, and FN (false negative) represents the number of non-landslide samples incorrectly classified. P and N denote the total number of positive and negative samples, respectively.

4. Result

4.1. Exploration and Selection of Landslide Conditioning Factors

GeoDetector can handle both numerical variables (e.g., slope, elevation, NDVI) and qualitative variables (e.g., lithology, land use). However, for continuous data, reclassification is necessary before applying GeoDetector to enhance spatial analysis [55]. Previous studies employing GeoDetector for LCFs selection typically used manual or natural breaks methods to categorize continuous variables, with no standardized approach for determining the number of classification intervals [18,25]. However, for a spatial statistical model like factor GeoDetector, the classification of variables can significantly influence the calculation of the explanatory power, represented by the q value. In this study, five methods: equal interval, natural breaks, quantile, geometric interval, and standard deviation, were employed to determine the optimal reclassification scheme for each continuous variable. Considering the data volume and computational complexity, the classification number range for all methods was set to the commonly used [3, 9]. The highest q value for each factor across different classification methods and numbers of intervals was identified as the factor's optimal explanatory power for landslides. The variation in q values for each LCF under different reclassification schemes is shown in Figure 7. For example, for elevation, the best explanation of its influence on landslides is achieved when it is classified into six classes using the natural breaks method.

Ultimately, GeoDetector was applied based on the optimal reclassification scheme for each factor, and the factor detection results for all LCFs are shown in Figure 8. It can be observed that precipitation and elevation have the strongest potential influence on landslides in the study area, which is consistent with our previous findings using the GWR model to investigate landslide driving factors in Luhe [56]. According to statistics, the average elevation of 166 landslide points is only 200.87 m, often located in low-altitude areas with frequent human activity. As elevation decreases, the intensity of road construction and land development increases. The q values of DTR and land use rank third and fourth, respectively, which not only validates the reasonableness of the GeoDetector results but

also highlights that, in addition to precipitation, anthropogenic factors are a significant cause of landslides in the study area.

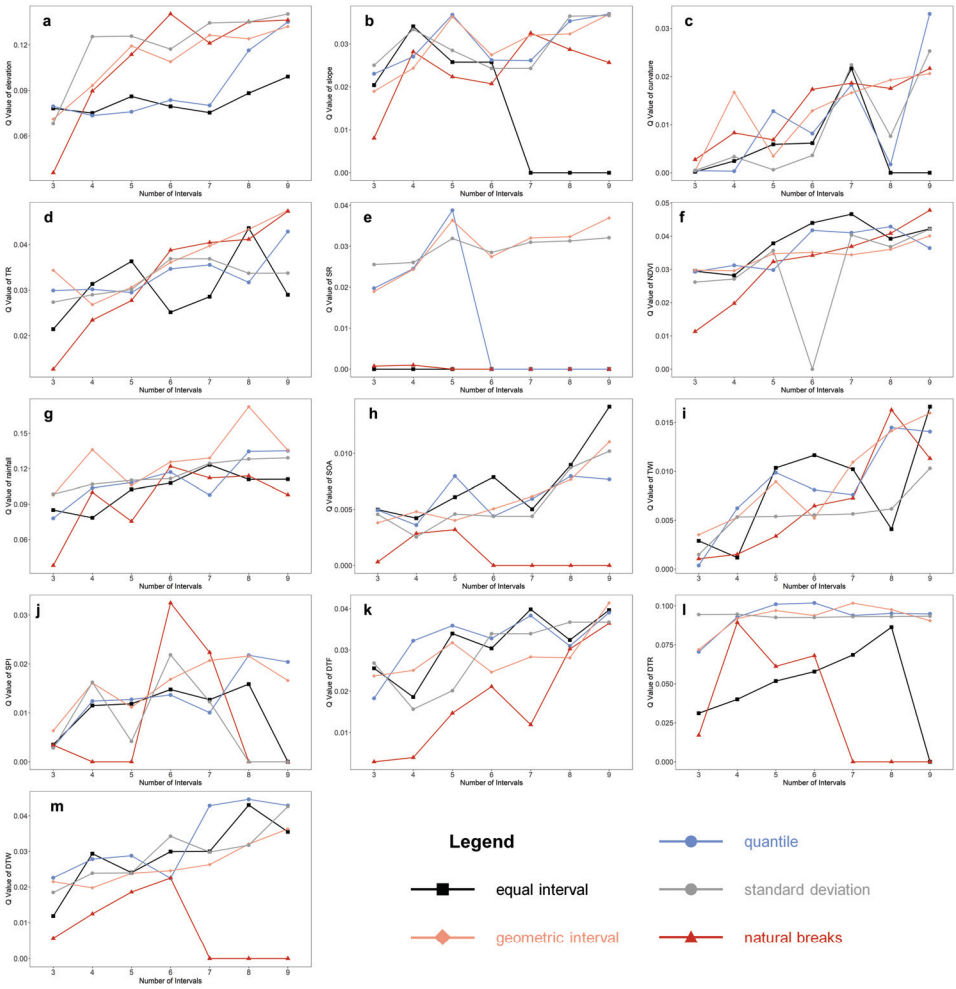


Figure 7. Optimal reclassification results for continuous variables. The *x*-axis represents the number of classes, and the *y*-axis represents the q value. (**a**) elevation; (**b**) slope; (**c**) curvature; (**d**) TR; (**e**) SR; (**f**) NDVI; (**g**) precipitation; (**h**) SOA; (**i**) TWI; (**j**) SPI; (**k**) DTF; (**l**) DTR; (**m**) DTR.

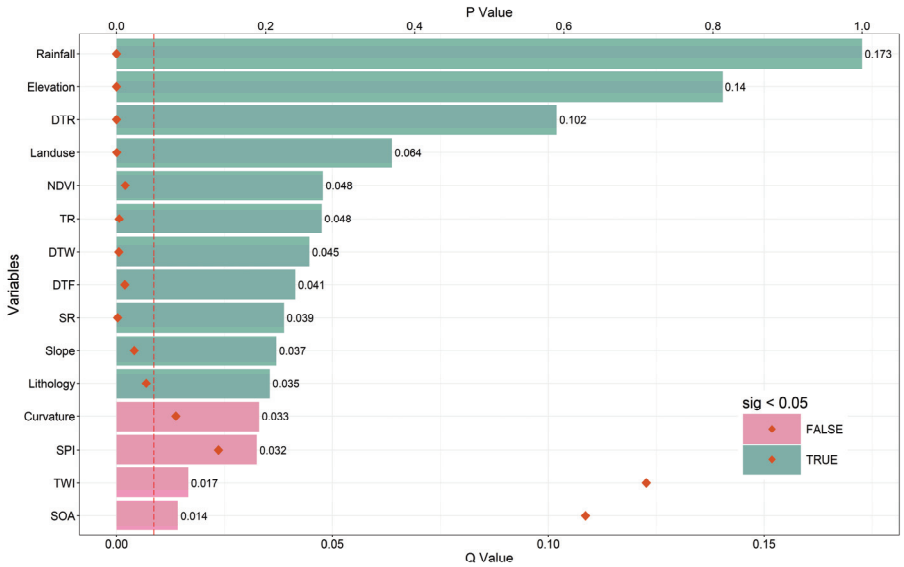


Figure 8. The optimal q statistic index and *p* values of LCFs obtained through GeoDetector.

The p values corresponding to the q values of the LCFs indicate the significance of each influencing factor, with the confidence interval for the significance test set at 95%. As observed, the factors curvature, SPI, TWI, and SOA did not pass the significance test, and their q values are relatively low. This suggests that these four factors have a weaker explanatory power for landslide spatial distribution. As a result, we excluded these four factors and selected the remaining 11 factors: elevation, slope, terrain relief, surface roughness, normalized difference vegetation index, precipitation, distance to fault, distance to road, distance to water, lithology, and land use, as the main factors for further analysis in subsequent models.

4.2. Generation of Non-Landslide Samples Based on the Information Value Model

The selected landslide conditioning factors were analyzed using GIS, with continuous variables classified using the natural breaks method. For each factor, the number of landslides and the area within each classified interval were extracted. The factors DTF, DTR, and DTW were classified based on empirically determined influence ranges. Subsequently, the proportions of landslide occurrences and area coverage were calculated for each interval. For intervals containing no landslide points, an average information value was considered, and an information value of –2 was directly assigned. The resulting information values for each factor, calculated using Equation (1), are presented in Table 2.

Table 2. Information values of landslide conditioning factors.

Conditioning Factors	Classes	Percentage of Area (%)	Percentage of Landslides (%)	IV
Elevation (m)	<159	35.534	57.229	0.477
	160–329	23.684	22.892	−0.034
	329–508	17.804	13.855	−0.251
	508–702	15.457	4.819	−1.165
	>702	7.521	1.205	−1.831
Slope (°)	<8	25.495	34.337	0.298
	8–16	30.539	39.759	0.264
	16–24	23.339	17.470	−0.290
	24–33	14.849	7.831	−0.640
	>33	5.779	0.602	−2.261
Terrain relief (m)	<17	22.682	36.747	0.482
	17–31	31.038	36.145	0.152
	31–45	26.149	20.482	−0.244
	45–63	15.332	4.217	−1.291
	>63	4.800	2.410	−0.689
Surface roughness	<1.04	58.636	74.699	0.242
	1.04–1.13	28.385	21.084	−0.297
	1.13–1.26	9.776	3.614	−0.995
	1.26–1.49	2.837	0.602	−1.550
	>1.49	0.367	0.000	−2.000
NDVI	<0.13	6.223	8.434	0.304
	0.13–0.22	13.906	26.506	0.645
	0.22–0.28	21.117	22.892	0.081
	0.28–0.35	35.430	32.530	−0.085
	>0.35	23.324	9.639	−0.884

Table 2. *Cont.*

Conditioning Factors	Classes	Percentage of Area (%)	Percentage of Landslides (%)	IV
Precipitation (mm)	<1603	39.058	61.446	0.453
	1603–1622	21.062	19.277	−0.089
	1622–1727	17.615	7.229	−0.891
	1727–1801	13.904	8.434	−0.500
	>1801	8.361	3.614	−0.839
Distance to fault (m)	<500	8.743	9.639	0.097
	500–1000	9.368	10.241	0.089
	1000–1500	8.361	13.253	0.461
	1500–2000	6.872	4.819	−0.355
	>2000	66.655	62.048	−0.072
Distance to road (m)	<50	19.159	40.964	0.760
	50–100	14.258	24.699	0.549
	100–150	11.801	11.446	−0.031
	150–200	9.862	5.422	−0.598
	>200	44.921	17.470	−0.944
Distance to water (m)	<50	11.182	21.084	0.634
	50–100	10.316	15.663	0.418
	100–150	9.911	16.265	0.495
	150–200	7.234	10.241	0.348
	>200	61.357	36.747	−0.513
Lithology	massive intrusive rock	80.806	86.747	0.071
	stratified clastic rock	14.165	7.831	−0.593
	single layer soil	4.232	4.217	−0.004
	multilayer soil	0.346	0.000	−2.000
	double layer soil	0.451	1.205	0.982
Land use	arable land	17.429	45.783	0.966
	woodland and shrubland	76.730	43.976	−0.557
	grassland	1.946	0.602	−1.173
	artificial surfaces	3.304	8.434	0.937
	water bodies and bare land	0.591	1.205	0.712

Using the raster calculator in GIS, the factor layers were overlaid based on Equation (2) to generate a preliminary landslide susceptibility map for Luhe County. This map was then classified into five susceptibility levels using the natural breaks method (Figure 9a). Very low and low susceptibility zones identified by the information value model were designated as non-landslide areas. Within these zones, a number of non-landslide points equal to twice the number of landslide samples were randomly generated. Additionally, it was ensured that all non-landslide points were located at least 500 m away from any positive sample (Figure 9b). This strategy aims to (1) prevent the LCFs attributes of non-landslide samples from being too similar to those of positive samples or to one another, which could otherwise affect model training, and (2) increase the number of negative samples—while maintaining their reliability—to reduce the risk of overfitting when the overall sample size is limited. Ultimately, 70% of all samples were used for model training, while the remaining 30% were reserved for validation.

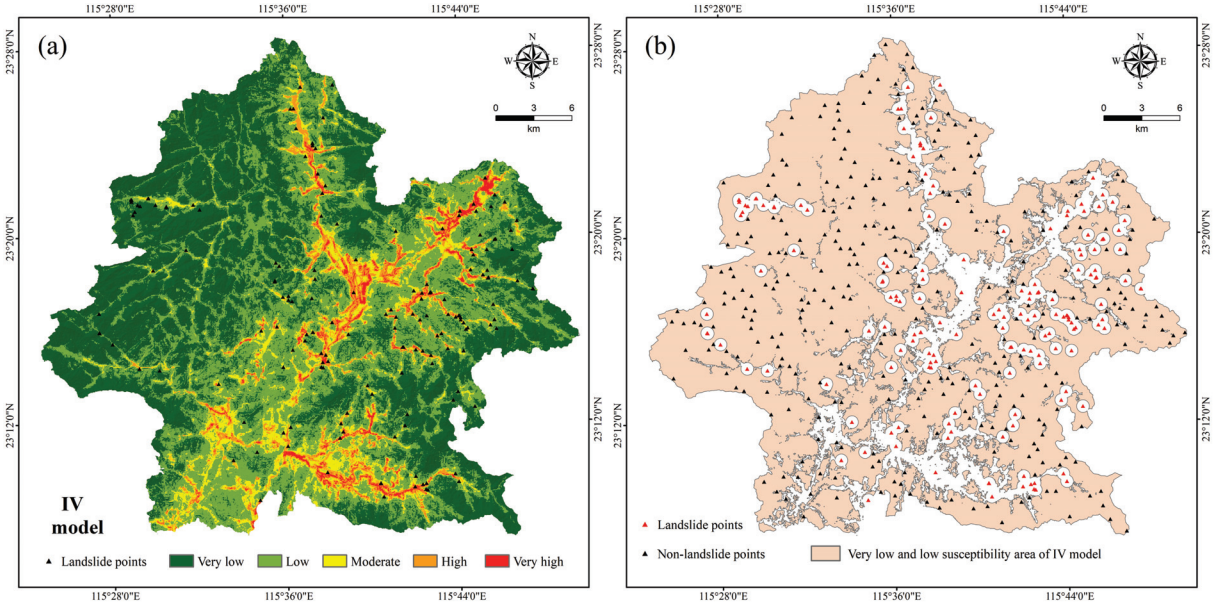


Figure 9. (**a**) Landslide susceptibility zonation based on the information value model. (**b**) Sample selection results within very low and low susceptibility zones.

4.3. Model Implementation and Performance Evaluation

In order to investigate the impact of GeoDetector factor selection on the model and compare the performance differences between local and global machine learning models, four models were constructed: RF, GeoD-RF, GWRF, and GeoD-GWRF. The initial factors and the selected factors were used as inputs for these models, respectively. To avoid overfitting in the GWRF model, grid search was employed to determine the optimal combination of hyperparameters. For each step size, 5-fold cross-validation was used to mitigate the risk of random variations in the parameters. The final hyperparameter settings for GWRF were as follows: adaptive kernel, bandwidth = 80, ntree = 500, mtry = 4. For uniform comparison, the ntree and mtry parameters for the RF model were kept consistent.

The landslide susceptibility evaluation was conducted using the test dataset for the study area, and the prediction results were compared with the landslide validation dataset. The ROC curves and corresponding AUC values for different models are shown in Figure 10. As can be seen, the performance of the local machine learning model (GWRF), which considers spatial structure, is significantly better than that of the global machine learning model (RF). The AUC value increased from 0.915 to 0.937, a 2.41% improvement. This indicates that GWRF, by considering the spatial heterogeneity of the LCFs, better captured the nonlinear relationship between LCFs and landslides at the regional scale, taking into account the local importance of LCFs. Furthermore, after factor selection using GeoDetector, although the number of factors decreased from 15 to 11, the model performance did not deteriorate. On the contrary, both GeoD-RF and GeoD-GWRF were able to achieve better prediction results with a reduced number of factors. This demonstrates that the factor optimization by GeoDetector is reliable. It not only helps eliminate unimportant redundant data, thereby reducing the burden on the model, but also further enhances the accuracy of the model's predictions.

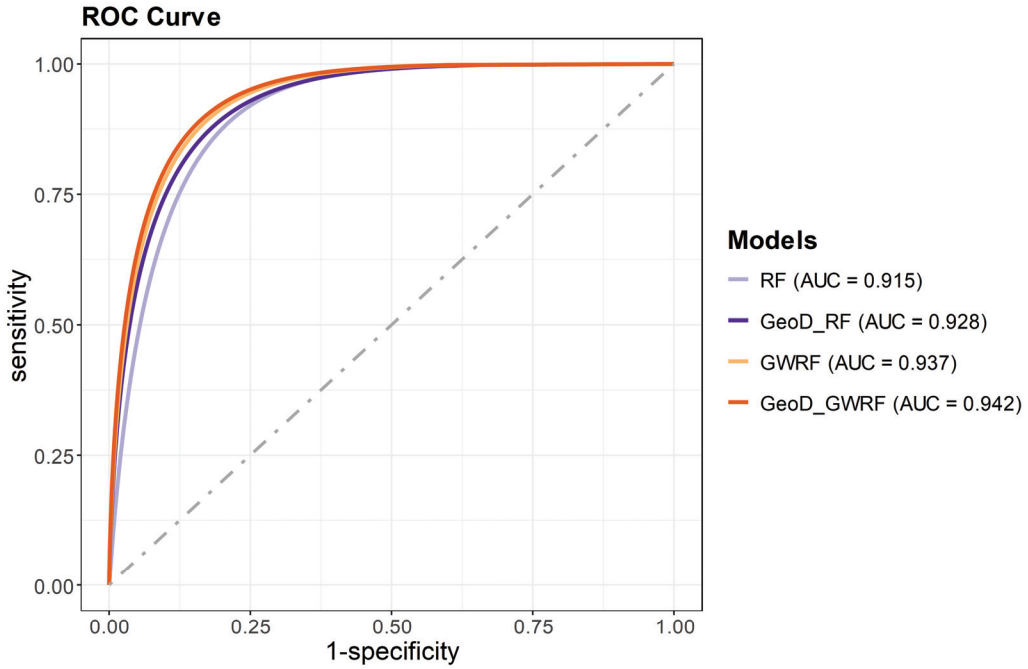


Figure 10. ROC curves for the four models with test data and their corresponding AUC values.

In addition, we conducted a spatial autocorrelation test on the squared residuals of the predictions made by the four models, using Moran's I as the test metric (Table 3). The range of Moran's I is between −1 and 1. A positive value indicates spatial clustering of the elements, a negative value indicates spatial dispersion, and values close to 0 suggest spatial randomness [57]. As shown, all four models exhibited a slight spatial clustering effect in their prediction residuals. However, compared to RF, the GWRF model had a smaller Moran's I value, indicating that its residuals were more spatially random. This suggests that the local decomposition method, which accounts for the spatial non-stationarity, is less affected by the spatial heterogeneity of the factors themselves and is therefore suitable for broader study areas. It is worth emphasizing that after factor optimization using GeoDetector, the Moran's I values further decreased, with GeoD-GWRF achieving the smallest Moran's I value of 0.106. This indicates that, after optimization by GeoDetector, the impact of LCFs' spatial heterogeneity on the prediction results was further reduced. The factors excluded by GeoDetector were those with weaker spatial effects on landslides. These factors exerted a smaller driving influence on landslides, and their inclusion would have increased the model's prediction uncertainty across the entire study area. This further validates the applicability and rationality of the GeoDetector.

Table 3. Spatial autocorrelation test of the squared prediction residuals for the four models.

Model	Moran's I	z-Score	*p*-Value
RF	0.157	6.132	0.000
GeoD-RF	0.129	4.643	0.000
GWRF	0.137	5.498	0.000
GeoD-GWRF	0.106	4.103	0.000

4.4. Differences in Landslide Susceptibility Mapping Across Models

The dataset for all grid cells was input into the four trained models to obtain landslide susceptibility evaluation results for the entire study area. The commonly used natural breaks method [24,58] was applied to classify each susceptibility layer into five susceptibility zones: very low, low, moderate, high, and very high. Finally, the landslide susceptibility

maps for the four models are shown in Figure 11. By overlaying the landslide distribution map of the study area, it can be observed that the high and very high susceptibility zones from the four models are strongly correlated with the distribution of landslides.

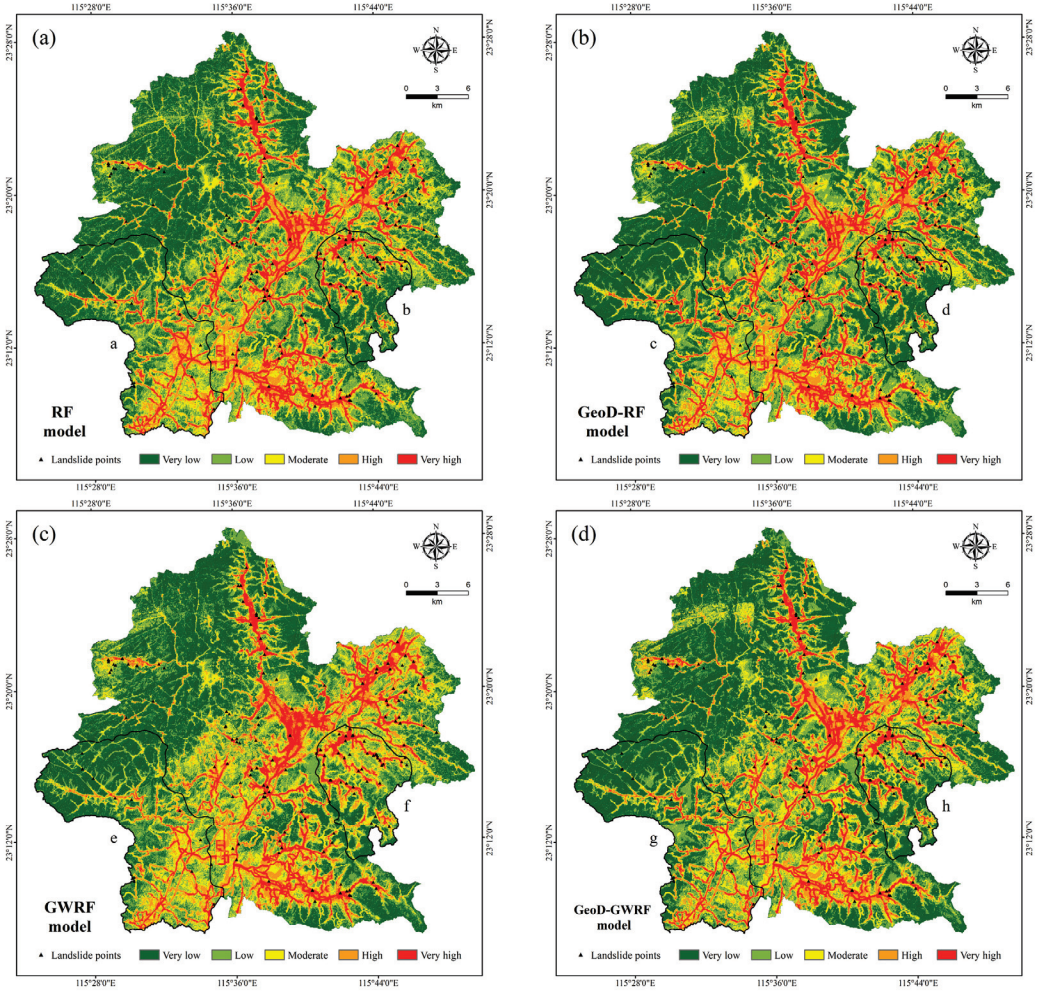


Figure 11. Landslide susceptibility zoning maps obtained from the four models: (**a**) RF model, (**b**) GeoD-RF model, (**c**) GWRF model, (**d**) GeoD-GWRF model.

Further statistical analysis of the results for each zone in relation to the landslide distribution (Figure 12) reveals that the area proportion of each zone decreases gradually from very low to very high susceptibility. However, the number of landslides in each zone increases progressively. This confirms the validity of the predictions made by the four models. The areas classified as very low susceptibility zones are larger in the GeoD-RF and GeoD-GWRF models (36.99% and 36.62%, respectively), while the RF model only classifies 33.64% of the area as very low susceptibility. At the same time, the very high susceptibility zones defined by GWRF and GeoD-GWRF are smaller (9.60% and 9.59%, respectively), but they predict more landslides in these zones (88 and 91, respectively). The local machine learning methods classify more areas into low susceptibility zones and fewer areas into high susceptibility zones, yet they yield more accurate prediction results. This is because local learning models account for the varying influence of factors in different regions, overcoming the limitation of global models, where factor importance is assumed to be constant across space. This makes the classification results more targeted. It is noteworthy that, compared to RF and GWRF, the GeoD-RF and GeoD-GWRF models, after factor optimization, predict more landslides with less area classified as very high susceptibility. Although the GeoD-GWRF model uses the fewest LCFs, it produces the most accurate

zonal results. This indicates that the inclusion of redundant factors can actually reduce model performance, and the use of the GeoDetector method to exclude irrelevant LCFs is therefore justified.

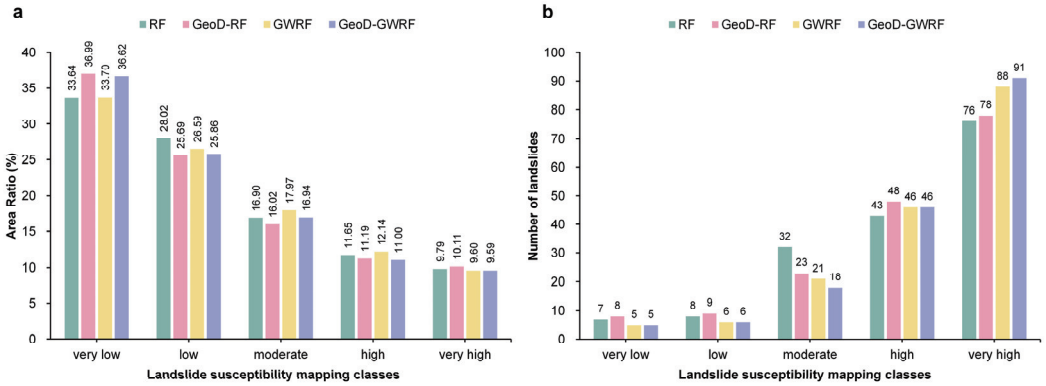


Figure 12. Statistical indicators of landslide susceptibility mapping. (**a**) Area proportion of each zone, (**b**) Number of landslides within each zone.

5. Discussion

5.1. Local Interpretability of the GWRF Model

As a local decomposition of the RF model, the GWRF model's results can be mapped individually, making it a useful exploratory tool [51]. Similar to the RF model, GWRF can also calculate the relative importance of factors to landslides by computing the OOB mean square error (Equation (6)). However, this importance is not spatially constant; instead, for each local model decomposition (corresponding to each sample point), there is a relative importance value. Specifically, the GWRF model allows factors to have different influence weights on landslides at different spatial locations, making the spatial influence of factors on landslides interpretable.

We averaged the LCFs' importance calculated by all local models and used this as the global importance of LCFs computed by GWRF, which was then compared with the RF model (Figure 13a). It should be noted that the comparison here is based on GeoD-RF and GeoD-GWRF models, which were optimized by GeoDetector. At the global scale, the two models show little difference in the factors they consider most strongly correlated with landslides. Both models identify distance to roads, elevation, distance to water, terrain relief, and precipitation as the most important factors. However, the GWRF model regards distance to faults as more strongly correlated with landslides than NDVI. This may be because the NDVI is generally high in Luhe (Figure 3f), and the characteristics of NDVI in most of the divided local areas are similar, which weakens its influence. It is worth discussing that GeoDetector assesses the driving effect of factors on landslides by examining the spatial distribution of factors and their similarity to the spatial pattern of landslides across the entire study area, based on the spatial heterogeneity of the factors themselves. GWRF, on the other hand, decomposes the study area through bandwidths and calculates the relative importance by performing random forest simulations on each bandwidth. While this approach also takes into account the spatial heterogeneity of factors to some extent, the results inevitably differ from those obtained by GeoDetector. Interestingly, however, the results from RF and GWRF still show a high degree of similarity with the factor detection results from GeoDetector (Figure 8). All three methods identify elevation, precipitation, and distance to roads as having a significant impact on landslides.

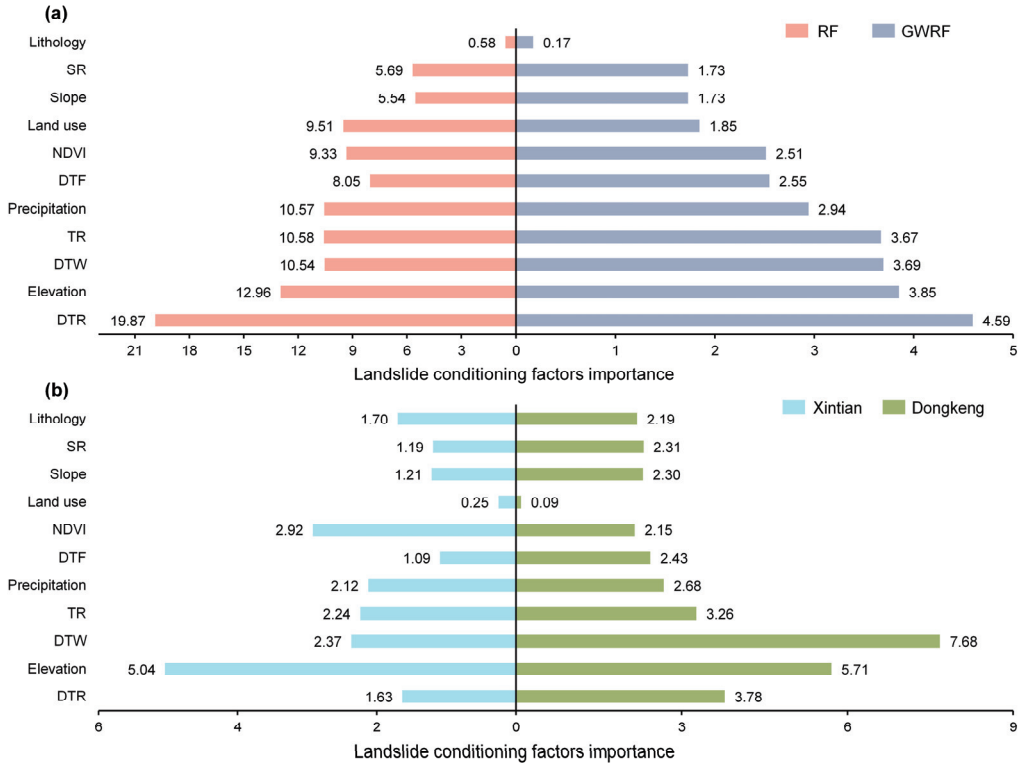


Figure 13. Importance of LCFs obtained by each model. (**a**) Global importance of LCFs obtained by the RF and GWRF models. (**b**) Importance of LCFs obtained by the GWRF model in Xintian and Dongkeng, respectively.

Although the factor importance results from GWRF and RF are generally similar, there are differences in the factor importance results obtained by each local model within the GWRF framework. The characteristic of local weight reallocation in GWRF also leads to spatial differences in the susceptibility mapping results. To better illustrate the differences between local and global model predictions, we further extracted the susceptibility prediction results of the four models in Xintian, located in the southwestern part with fewer disaster points, and Dongkeng, in the eastern part with more disaster points (Figure 14). Additionally, we calculated the factor importance derived from GWRF in these two regions (Figure 13b).

As shown, in Xintian, which has fewer disaster points, RF and Geo-RF still predict a larger portion of the area as high susceptibility zones (Figure 14a,c). In contrast, GWRF and Geo-GWRF predict most of Xintian as low susceptibility zones (Figure 14e,g). As a global model, RF still considers DTR and elevation to be the main factors influencing landslides in Xintian (Figure 13a). However, GWRF identifies elevation and NDVI as the primary factors affecting landslides in this region (Figure 13b). The change in factor importance leads to differences in the landslide susceptibility predictions between the two models for this region. Although both models accurately predict all the landslides in Xintian, the GWRF model predicts a smaller area for high susceptibility zones. In other words, the GWRF prediction is more targeted because it identifies the most important factors influencing landslides in this region. For Dongkeng, which has more disaster points, GWRF identifies the five most important factors influencing landslides as DTW, elevation, DTR, TR, and precipitation (Figure 13b), while RF still considers DTR, elevation, TR, precipitation, and DTW to be the most important factors in sequence (Figure 13a). Since the most important factors remain largely unchanged, the prediction results from both models are quite similar, and each model makes correct predictions (Figure 14d,h). However, the GWRF model assigns higher weight proportions to DTW in this region, which leads to more areas

near rivers being classified as high susceptibility zones. In summary, the local learning characteristics of the GWRF model enable it to better capture the impact of LCFs' spatial heterogeneity on landslides. By considering the differences in the importance of the same factor across different regions, it makes more targeted judgments on the final landslide susceptibility predictions. This not only improves the accuracy of the predictions but also helps reduce the associated costs of disaster prevention and mitigation efforts.

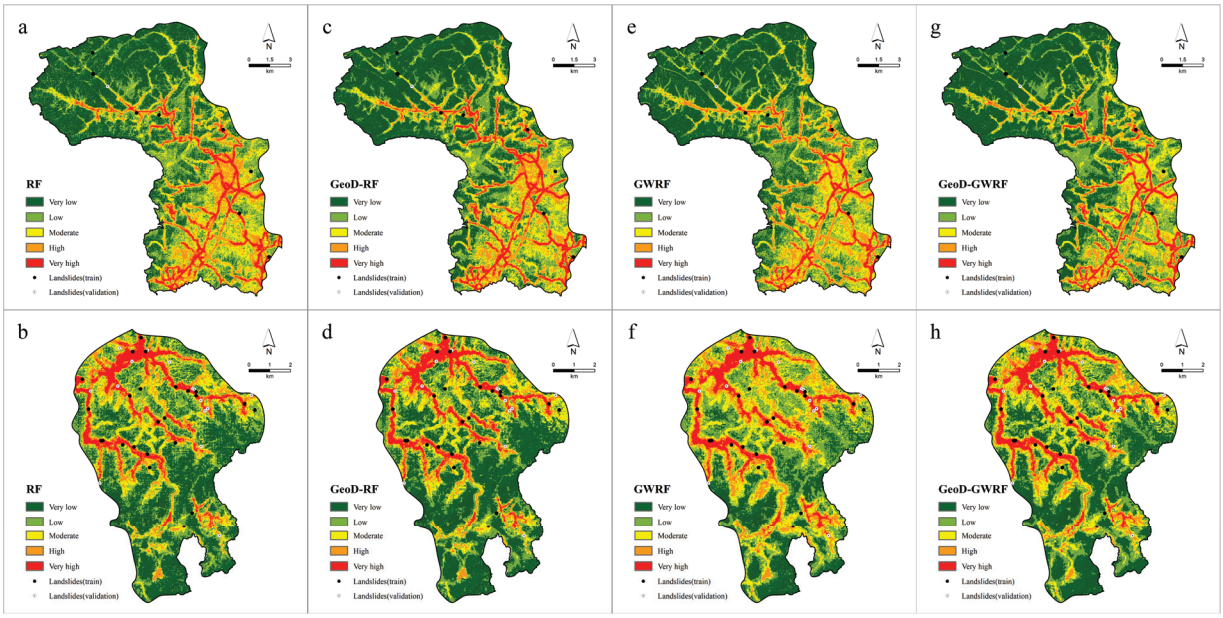


Figure 14. Prediction differences of each model in different regions. Panels (**a**) to (**g**) correspond to the labeled regions in Figure 11. Panels (**a**), (**c**), (**e**), and (**g**) show the predictions of the four models in Xintian, while panels (**b**), (**d**), (**f**), and (**h**) display the predictions of the four models in Dongkeng.

Based on the local interpretability of the GWRF model, we compiled the results from all local models within each town and identified the two most influential factors affecting landslides in each area (Figure 15). It can be observed that elevation and distance to roads play significant roles in multiple towns. In Dongkeng and Shuichun—areas severely affected by landslides—distance to roads ranked as the second most important factor, indicating that elevation-driven human activities, especially road construction, have a strong impact on landslide occurrence in these regions. In Hetian, river-induced slope disturbances emerged as the primary trigger for landslides. Meanwhile, in Nanwan and Shanghu, the dominant influencing factors were distance to faults and lithology, respectively, suggesting that greater attention should be paid to fault activity and lithological variation in these areas when assessing landslide risk.

To better account for the spatial heterogeneity of landslide conditioning factors, we further examined the influence of each factor across different geomorphic types. The geomorphic classification was based on the macro-geomorphological system of China proposed by Chenghu et al. (2009) [59], which utilizes the terrain relief index to divide landforms into five types: plain, platform, hill, low relief mountain, middle relief mountain, high relief mountain, and highest relief mountain. Referring to the latest "China land 1:1 million digital geomorphic classification system" [60,61] and considering the scale of the study area, we simplified the classification into four geomorphic types: plain (<10 m), platform (10–30 m), hilly (30–50 m), and low mountain (>50 m) (Figure 16a). We then identified the dominant LCF within each geomorphic type (Figure 16b). The plain areas are primarily located in the low-elevation zones of the study area, where elevation emerged as the most important influencing factor. Notably, the average elevation of all

166 landslide points was only 200.87 m, significantly lower than the study area's mean elevation of 314 m. This indicates that landslides are more likely to occur in low-lying zones with intensive human activity, a trend consistent with findings from many studies in southeastern coastal China [7,62,63]. It suggests that anthropogenic factors—such as infrastructure development and land use change—may be the primary drivers of local hazards in these plain regions. In the platform and hilly areas, the dominant LCFs were distance to water and distance to road, respectively. This implies that in platform zones, increasing terrain variation enhances the erosive and scouring effects of water on slopes, making hydrological processes a major driver of landslides. In contrast, in hilly regions, road construction poses a greater disturbance to slope stability, indicating the need for proper road planning, as well as reinforcement of roadside slopes and drainage systems. Low mountain areas, characterized by high elevation and significant terrain variation, tend to facilitate the development of vertical joints. As elevation increases, gravitational potential energy acting on slopes also rises, making these areas more susceptible to landslides. Here, terrain relief was identified as the most critical factor influencing landslide occurrence. By visualizing the local importance of each factor in both township and geomorphic units, the GWRF model provides new insights that can guide region-specific disaster prevention and mitigation efforts.

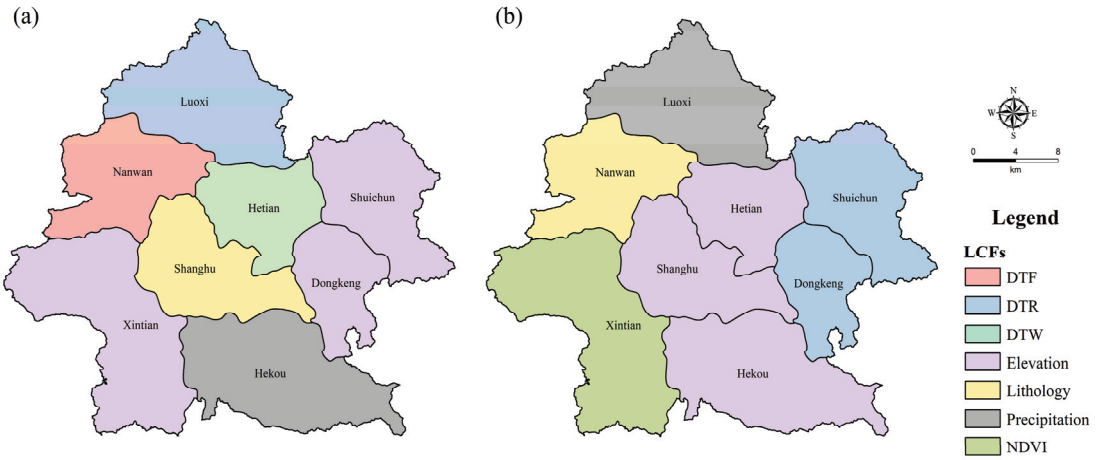


Figure 15. Key landslide conditioning factors in each township. (**a**) Primary key LCF in each township, (**b**) Secondary key LCF in each township.

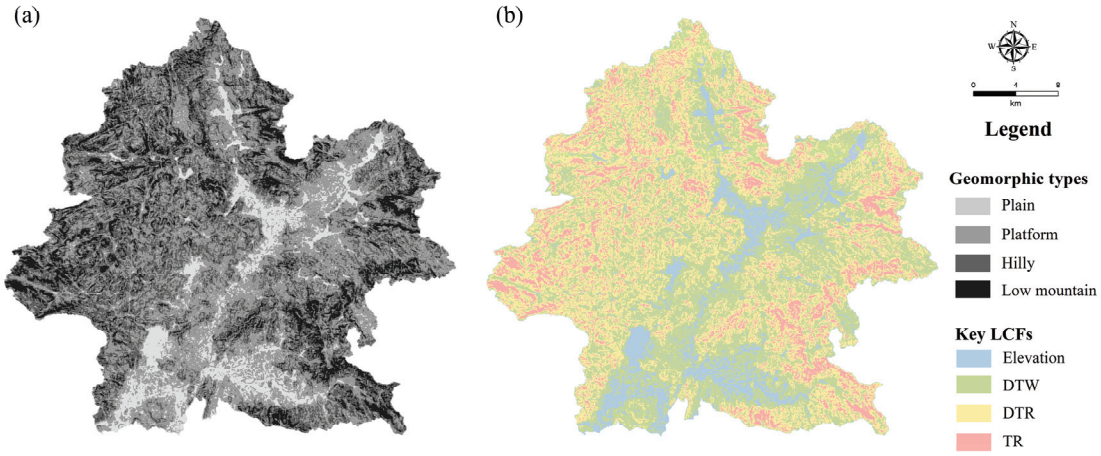


Figure 16. Key landslide conditioning factors in each geomorphic type. (**a**) Geomorphic distribution of the study area. (**b**) Dominant LCFs in each geomorphic type.

5.2. Generalizability of the Proposed Method for Landslide Susceptibility Mapping

To enhance the accuracy of landslide susceptibility assessment, this study employed the *p*-value threshold from GeoDetector's significance test to select the LCFs most strongly associated with landslide occurrences. The selected factors were then analyzed using the GWRF model, which incorporates local modeling to account for spatial heterogeneity. To evaluate the effectiveness of the proposed approach, we compared it with two widely used machine learning models: logistic regression [11] and support vector machine (SVM) [14]. Additionally, we applied multicollinearity testing as an alternative method for LCF selection, and compared its performance against GeoDetector-based selection when combined with the different models. Multicollinearity among variables was assessed using two common indicators: tolerance (TOL) and the variance inflation factor (VIF). When TOL < 0.2 and VIF > 5, multicollinearity is considered to be significant among the explanatory variables [37]. A high VIF indicates a strong correlation among factors, suggesting potential data redundancy. As shown in Table 4, the initial multicollinearity test revealed strong collinearity between elevation and precipitation, as well as between slope and surface roughness. Considering this, the surface roughness and precipitation factors were excluded. After removal, the remaining factors exhibited only weak correlations (VIF < 5), indicating acceptable levels of multicollinearity.

Table 4. Multicollinearity analysis of LCFs.

Landslide Conditioning Factors	First Round of Analysis		Second Round of Analysis	
	TOL	**VIF**	**TOL**	**VIF**
Elevation	0.06	15.94	0.74	1.35
Slope	0.09	10.88	0.30	3.37
Curvature	0.83	1.20	0.86	1.16
Terrain relief	0.31	3.24	0.31	3.20
Surface roughness	0.13	7.77	-	-
NDVI	0.83	1.20	0.85	1.18
Precipitation	0.07	14.58	-	-
Slope of aspect	0.81	1.24	0.82	1.22
TWI	0.64	1.55	0.70	1.42
SPI	0.74	1.36	0.76	1.32
Distance to fault	0.90	1.11	0.91	1.10
Distance to road	0.85	1.18	0.87	1.16
Distance to water	0.82	1.22	0.88	1.14
Lithology	0.95	1.05	0.95	1.05
Land use	0.94	1.06	0.94	1.06

Subsequently, all factors, the factors selected through VIF, and the factors selected through GeoDetector were incorporated into each model. To examine potential overfitting issues, we applied a 5-fold cross-validation method, randomly dividing the sample set for five rounds of training and validation. The median and standard deviation of accuracy, AUC, recall, and F1 score for each model were calculated (Table 5), and ROC curves for each model were plotted (Figure 17). The results showed that the performance of each model across different data subsets was stable, with small standard deviations, indicating good generalization capability and a low risk of overfitting. After factor selection using VIF, both the accuracy and AUC values for the logistic regression model improved, and the model's recall and F1 score were 0.580 and 0.611, respectively, surpassing the Geo-Logistic model. However, for the SVM, RF, and GWRF models, the performance metrics slightly declined. This suggests that while the VIF method helped reduce data redundancy and multicollinearity to some extent, it also removed some factors strongly related to landslides,

leading to a decrease in model accuracy. Additionally, multicollinearity issues tend to affect linear regression models like logistic regression more significantly, while machine learning models are less influenced by correlations between variables and are more focused on the relationship between dependent and independent variables. For the SVM model, the AUC values were 0.843, 0.842, and 0.844 for the three factor sets, demonstrating that the SVM model showed good adaptability to different data and exhibited better generalization ability and stability. In contrast, the RF and GWRF models, based on binary tree splitting mechanisms, have higher data quality requirements, and the results varied considerably across different factor sets. After removing the rainfall factor, which had a strong influence on landslides, using VIF, the performance of both models significantly decreased. Overall, the dataset constructed using the factors selected by GeoDetector's q and p threshold values effectively captured the spatial correlation between the dependent and independent variables. It removed less influential factors, reducing data size while enhancing the performance of the machine learning models. Based on comparisons of accuracy and AUC, the GWRF model consistently demonstrated superior performance. This local modeling approach accounts for geographic, topographic, and environmental factors in the region, making it more effective for predicting and modeling small spatial clusters, resulting in more accurate and targeted predictions.

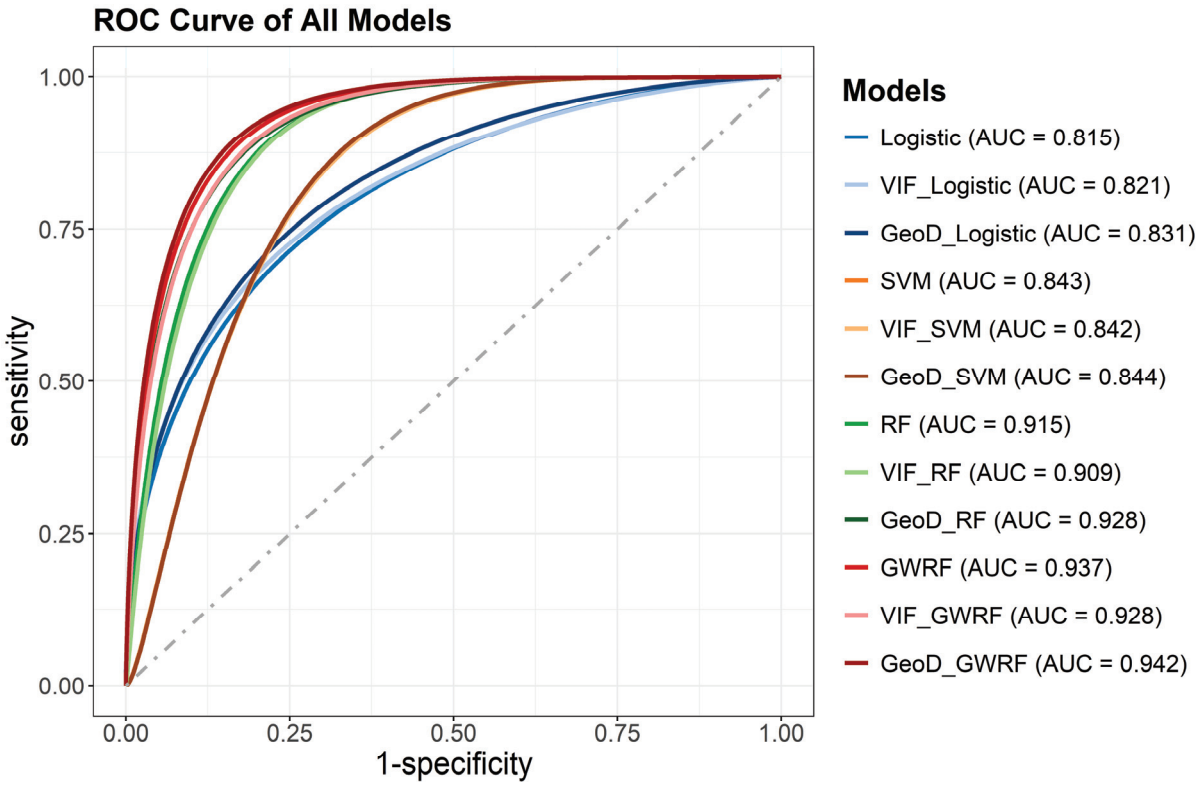


Figure 17. ROC curves for the 12 models with test data and their corresponding AUC values.

Table 5. Performance Metrics Statistics for Each Model.

Model	Accuracy		AUC		Recall		F1 Score	
	Median	Std	Median	Std	Median	Std	Median	Std
Logistic	0.740	0.012	0.815	0.008	0.560	0.016	0.589	0.012
VIF-Logistic	0.753	0.016	0.821	0.015	0.580	0.017	0.611	0.013
GeoD-Logistic	0.753	0.009	0.831	0.006	0.560	0.008	0.602	0.013
SVM	0.833	0.009	0.843	0.017	0.580	0.007	0.699	0.007
VIF-SVM	0.820	0.024	0.842	0.016	0.560	0.016	0.675	0.018

Table 5. *Cont.*

Model	Accuracy		AUC		Recall		F1 Score	
	Median	**Std**	**Median**	**Std**	**Median**	**Std**	**Median**	**Std**
GeoD-SVM	0.833	0.025	0.844	0.010	0.600	0.018	0.706	0.016
RF	0.840	0.016	0.915	0.024	0.640	0.019	0.727	0.009
VIF-RF	0.833	0.017	0.909	0.024	0.620	0.018	0.713	0.016
GeoD-RF	0.853	0.017	0.928	0.015	0.680	0.016	0.756	0.018
GWRF	0.880	0.020	0.937	0.030	0.780	0.016	0.812	0.019
VIF-GWRF	0.867	0.017	0.928	0.015	0.740	0.017	0.787	0.019
GeoD-GWRF	0.880	0.009	0.942	0.012	0.780	0.009	0.812	0.016

5.3. Limitations of the Current Study

Although the Geo-GWRF model effectively removes redundant factors unrelated to landslides and addresses the linear limitations and inefficiency issues faced by traditional geographically weighted regression when dealing with large spatial datasets [64], there are still some limitations in this study that warrant further improvement in future work.

First, the criteria for factor selection using GeoDetector were somewhat limited. In this study, factors unrelated to landslides were removed based on a *p*-value threshold of 0.05. However, some studies have also achieved good results by directly defining q-values or combining other factor selection methods, such as information gain ratio [25,65]. The standard for factor removal when using GeoDetector for LSM has not yet been standardized, and the most optimal method still warrants further discussion in future research. Additionally, Luhe is located in a subtropical monsoon climate zone, and the consideration of factors such as seasonal precipitation is still part of the future work plan.

Secondly, there are still some limitations in the interpretability of the GWRF model itself. GWRF calculates the strength of factor importance by assessing error increments, but due to the "black-box" nature of machine learning, it cannot reveal the direction of the factor's influence. In other words, we cannot determine whether an increase or decrease in a specific factor's characteristics leads to a positive or negative effect on landslides. However, we have noticed that a machine learning interpretation method called SHapley Additive exPlanations (SHAP) has shown promising results [66,67]. We aim to combine this method with GWRF to better understand the impact of LCFs on landslides at the local scale.

6. Conclusions

This study proposes a new method that combines GeoDetector and GWRF local machine learning to fully account for the impact of spatial heterogeneity of landslide conditioning factors in landslide susceptibility prediction. The results show that the GeoD-GWRF model can effectively remove disaster-related factors that are not associated with landslides from a spatial differentiation perspective. Based on the results from the ROC curve and Moran's I, the GeoD-GWRF model achieves higher prediction accuracy and is applicable to a broader study area. At the same time, the predictions from the local machine learning method are more targeted. By leveraging the local interpretability of GeoD-GWRF, related costs in disaster prevention and mitigation can be reduced, offering new perspectives for landslide prevention and control in different regions of the study area.

Author Contributions: Conceptualization, F.L. and G.Z.; Methodology, F.L. and G.Z.; Software, F.L., G.Z., T.W. and Y.Y.; Validation, F.L. and Q.Z.; Formal analysis, Y.Y. and Q.Z.; Resources, G.Z. and T.W.; Data curation, F.L., G.Z., T.W., Y.Y. and Q.Z.; Writing—original draft, F.L. and G.Z.; Writing—review and editing, F.L. and G.Z.; Visualization, F.L., G.Z., T.W. and Y.Y.; Supervision, Y.Y. and Q.Z.; Project administration, G.Z., T.W., Y.Y. and Q.Z.; Funding acquisition, G.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Guangdong Basic and Applied Basic Research Foundation (No. 2025A1515011669), the Guangdong Special Fund for National Park Construction (No. 2021GJGY026), and the Science and Technology Program of Guangzhou, China, under Grant (No. 201707010209).

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 authors would like to thank Junwei Zhen (Huizhou Geological Survey Center, Guangdong Provincial Bureau of Geology) for providing the original data used in this study. The authors are also grateful to the editor and the anonymous reviewers for their constructive comments and suggestions on the manuscript.

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

References

1. Han, Z.; Su, B.; Li, Y.; Ma, Y.; Wang, W.; Chen, G. Comprehensive analysis of landslide stability and related countermeasures: A case study of the Lanmuxi landslide in China. *Sci. Rep.* **2019**, *9*, 12407. [CrossRef]
2. Petley, D. Global patterns of loss of life from landslides. *Geology* **2012**, *40*, 927–930. [CrossRef]
3. Froude, M.J.; Petley, D.N. Global fatal landslide occurrence from 2004 to 2016. *Nat. Hazards Earth Syst. Sci.* **2018**, *18*, 2161–2181. [CrossRef]
4. Huang, F.; Zhang, J.; Zhou, C.; Wang, Y.; Huang, J.; Zhu, L. A deep learning algorithm using a fully connected sparse autoencoder neural network for landslide susceptibility prediction. *Landslides* **2020**, *17*, 217–229. [CrossRef]
5. Kaur, R.; Gupta, V.; Chaudhary, B. Landslide susceptibility mapping and sensitivity analysis using various machine learning models: A case study of Beas valley, Indian Himalaya. *Bull. Eng. Geol. Environ.* **2024**, *83*, 228. [CrossRef]
6. Panchal, S.; Shrivastava, A.K. Landslide hazard assessment using analytic hierarchy process (AHP): A case study of National Highway 5 in India. *Ain Shams Eng. J.* **2022**, *13*, 101626. [CrossRef]
7. Zhang, G.; Cai, Y.; Zheng, Z.; Zhen, J.; Liu, Y.; Huang, K. Integration of the statistical index method and the analytic hierarchy process technique for the assessment of landslide susceptibility in Huizhou, China. *Catena* **2016**, *142*, 233–244. [CrossRef]
8. Zhang, W.; Liu, S.; Wang, L.; Samui, P.; Chwała, M.; He, Y. Landslide susceptibility research combining qualitative analysis and quantitative evaluation: A case study of Yunyang County in Chongqing, China. *Forests* **2022**, *13*, 1055. [CrossRef]
9. Wang, Q.; Guo, Y.; Li, W.; He, J.; Wu, Z. Predictive modeling of landslide hazards in Wen County, northwestern China based on information value, weights-of-evidence, and certainty factor. *Geomat. Nat. Hazards Risk* **2019**, *10*, 820–835. [CrossRef]
10. Aditian, A.; Kubota, T.; Shinohara, Y. Comparison of GIS-based landslide susceptibility models using frequency ratio, logistic regression, and artificial neural network in a tertiary region of Ambon, Indonesia. *Geomorphology* **2018**, *318*, 101–111. [CrossRef]
11. Budimir, M.; Atkinson, P.; Lewis, H. A systematic review of landslide probability mapping using logistic regression. *Landslides* **2015**, *12*, 419–436. [CrossRef]
12. 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]
13. Yuan, R.; Chen, J. A novel method based on deep learning model for national-scale landslide hazard assessment. *Landslides* **2023**, *20*, 2379–2403. [CrossRef]
14. Huang, Y.; Zhao, L. Review on landslide susceptibility mapping using support vector machines. *Catena* **2018**, *165*, 520–529. [CrossRef]
15. Youssef, A.M.; Pourghasemi, H.R.; Pourtaghi, Z.S.; Al-Katheeri, M.M. Landslide susceptibility mapping using random forest, boosted regression tree, classification and regression tree, and general linear models and comparison of their performance at Wadi Tayyah Basin, Asir Region, Saudi Arabia. *Landslides* **2016**, *13*, 839–856. [CrossRef]
16. 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]

17. Sevgen, E.; Kocaman, S.; Nefeslioglu, H.A.; Gokceoglu, C. A Novel Performance Assessment Approach Using Photogrammetric Techniques for Landslide Susceptibility Mapping with Logistic Regression, ANN and Random Forest. *Sensors* **2019**, *19*, 3940. [CrossRef]
18. 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]
19. Sun, D.; Shi, S.; Wen, H.; Xu, J.; Zhou, X.; Wu, J. A hybrid optimization method of factor screening predicated on GeoDetector and Random Forest for Landslide Susceptibility Mapping. *Geomorphology* **2021**, *379*, 107623. [CrossRef]
20. Chalkias, C.; Polykretis, C.; Karymbalis, E.; Soldati, M.; Ghinoi, A.; Ferentinou, M. Exploring spatial non-stationarity in the relationships between landslide susceptibility and conditioning factors: A local modeling approach using geographically weighted regression. *Bull. Eng. Geol. Environ.* **2020**, *79*, 2799–2814. [CrossRef]
21. Erener, A.; Duzgun, H.S.B. Improvement of statistical landslide susceptibility mapping by using spatial and global regression methods in the case of More and Romsdal (Norway). *Landslides* **2010**, *7*, 55–68. [CrossRef]
22. Gu, T.; Li, J.; Wang, M.; Duan, P. Landslide susceptibility assessment in Zhenxiong County of China based on geographically weighted logistic regression model. *Geocarto Int.* **2022**, *37*, 4952–4973. [CrossRef]
23. Hong, H.; Pradhan, B.; Sameen, M.I.; Chen, W.; Xu, C. Spatial prediction of rotational landslide using geographically weighted regression, logistic regression, and support vector machine models in Xing Guo area (China). *Geomat. Nat. Hazards Risk* **2017**, *8*, 1997–2022. [CrossRef]
24. Zhao, Z.; Xu, Z.; Hu, C.; Wang, K.; Ding, X. Geographically weighted neural network considering spatial heterogeneity for landslide susceptibility mapping: A case study of Yichang City, China. *Catena* **2024**, *234*, 107590. [CrossRef]
25. Yang, J.; Song, C.; Yang, Y.; Xu, C.; Guo, F.; Xie, L. New method for landslide susceptibility mapping supported by spatial logistic regression and GeoDetector: A case study of Duwen Highway Basin, Sichuan Province, China. *Geomorphology* **2019**, *324*, 62–71. [CrossRef]
26. Ng, C.W.W.; Yang, B.; Liu, Z.Q.; Kwan, J.S.H.; Chen, L. Spatiotemporal modelling of rainfall-induced landslides using machine learning. *Landslides* **2021**, *18*, 2499–2514. [CrossRef]
27. Popescu, M.E. Landslide causal factors and landslide remediatial options. In Proceedings of the 3rd International Conference on Landslides, Slope Stability and Safety of Infra-Structures, Singapore, 11–12 July 2002; pp. 61–81.
28. 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]
29. Merghadi, A.; Yunus, A.P.; Dou, J.; Whiteley, J.; Binh, T.; Dieu Tien, B.; Avtar, R.; Abderrahmane, B. Machine learning methods for landslide susceptibility studies: A comparative overview of algorithm performance. *Earth-Sci. Rev.* **2020**, *207*, 103225. [CrossRef]
30. Zhang, T.; Han, L.; Chen, W.; Shahabi, H. Hybrid Integration Approach of Entropy with Logistic Regression and Support Vector Machine for Landslide Susceptibility Modeling. *Entropy* **2018**, *20*, 884. [CrossRef]
31. Li, L.-M.; Cheng, S.-K.; Wen, Z.-Z. Landslide prediction based on improved principal component analysis and mixed kernel function least squares support vector regression model. *J. Mt. Sci.* **2021**, *18*, 2130–2142. [CrossRef]
32. Chen, X.; Chen, W. GIS-based landslide susceptibility assessment using optimized hybrid machine learning methods. *Catena* **2021**, *196*, 104833. [CrossRef]
33. Wang, J.; Xu, C. Geodetector: Principle and prospective. *Acta Geogr. Sin.* **2017**, *72*, 116–134.
34. Wang, Y.; Wen, H.; Sun, D.; Li, Y. Quantitative Assessment of Landslide Risk Based on Susceptibility Mapping Using Random Forest and GeoDetector. *Remote Sens.* **2021**, *13*, 2625. [CrossRef]
35. Liu, Y.; Zhang, W.; Zhang, Z.; Xu, Q.; Li, W. Risk Factor Detection and Landslide Susceptibility Mapping Using Geo-Detector and Random Forest Models: The 2018 Hokkaido Eastern Iburi Earthquake. *Remote Sens.* **2021**, *13*, 1157. [CrossRef]
36. Quevedo, R.P.; Velastegui-Montoya, A.; Montalvan-Burbano, N.; Morante-Carballo, F.; Korup, O.; Renno, C.D. Land use and land cover as a conditioning factor in landslide susceptibility: A literature review. *Landslides* **2023**, *20*, 967–982. [CrossRef]
37. Ren, T.; Gao, L.; Gong, W. An ensemble of dynamic rainfall index and machine learning method for spatiotemporal landslide susceptibility modeling. *Landslides* **2024**, *21*, 257–273. [CrossRef]
38. Xiao, T.; Zhang, L.M.; Cheung, R.W.M.; Lacasse, S. Predicting spatio-temporal man-made slope failures induced by rainfall in Hong Kong using machine learning techniques. *Geotechnique* **2022**, *73*, 749–765. [CrossRef]
39. Zhu, A.X.; Miao, Y.; Wang, R.; Zhu, T.; Deng, Y.; Liu, J.; Yang, L.; Qin, C.-Z.; Hong, H. A comparative study of an expert knowledge-based model and two data-driven models for landslide susceptibility mapping. *Catena* **2018**, *166*, 317–327. [CrossRef]
40. Martinello, C.; Delchiaro, M.; Iacobucci, G.; Cappadonia, C.; Rotigliano, E.; Piacentini, D. Exploring the geomorphological adequacy of the landslide susceptibility maps: A test for different types of landslides in the Bidente river basin (northern Italy). *Catena* **2024**, *238*. [CrossRef]
41. Peng, L.; Niu, R.; Huang, B.; Wu, X.; Zhao, Y.; Ye, R. Landslide susceptibility mapping based on rough set theory and support vector machines: A case of the Three Gorges area, China. *Geomorphology* **2014**, *204*, 287–301. [CrossRef]

42. Nefeslioglu, H.A.; Gokceoglu, C.; Sonmez, H. An assessment on the use of logistic regression and artificial neural networks with different sampling strategies for the preparation of landslide susceptibility maps. *Eng. Geol.* **2008**, *97*, 171–191. [CrossRef]
43. Huang, F.; Yin, K.; Huang, J.; Gui, L.; Wang, P. Landslide susceptibility mapping based on self-organizing-map network and extreme learning machine. *Eng. Geol.* **2017**, *223*, 11–22. [CrossRef]
44. Khabiri, S.; Crawford, M.M.; Koch, H.J.; Haneberg, W.C.; Zhu, Y. An Assessment of Negative Samples and Model Structures in Landslide Susceptibility Characterization Based on Bayesian Network Models. *Remote Sens.* **2023**, *15*, 3200. [CrossRef]
45. Hu, Q.; Zhou, Y.; Wang, S.; Wang, F. Machine learning and fractal theory models for landslide susceptibility mapping: Case study from the Jinsha River Basin. *Geomorphology* **2020**, *351*, 106975. [CrossRef]
46. Dou, H.; He, J.; Huang, S.; Jian, W.; Guo, C. Influences of non-landslide sample selection strategies on landslide susceptibility mapping by machine learning. *Geomat. Nat. Hazards Risk* **2023**, *14*, 2285719. [CrossRef]
47. Wang, J.-F.; Li, X.-H.; Christakos, G.; Liao, Y.-L.; Zhang, T.; Gu, X.; Zheng, X.-Y. Geographical Detectors-Based Health Risk Assessment and its Application in the Neural Tube Defects Study of the Heshun Region, China. *Int. J. Geogr. Inf. Sci.* **2010**, *24*, 107–127. [CrossRef]
48. Luo, W.; Liu, C.-C. Innovative landslide susceptibility mapping supported by geomorphon and geographical detector methods. *Landslides* **2018**, *15*, 465–474. [CrossRef]
49. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
50. Sahin, E.K.; Colkesen, I.; Kavzoglu, T. A comparative assessment of canonical correlation forest, random forest, rotation forest and logistic regression methods for landslide susceptibility mapping. *Geocarto Int.* **2020**, *35*, 341–363. [CrossRef]
51. Georganos, S.; Grippa, T.; Gadiaga, A.N.; Linard, C.; Lennert, M.; Vanhuysse, S.; Mboga, N.; Wolff, E.; Kalogirou, S. Geographical random forests: A spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling. *Geocarto Int.* **2021**, *36*, 121–136. [CrossRef]
52. Dai, X.; Zhu, Y.; Sun, K.; Zou, Q.; Zhao, S.; Li, W.; Hu, L.; Wang, S. Examining the Spatially Varying Relationships between Landslide Susceptibility and Conditioning Factors Using a Geographical Random Forest Approach: A Case Study in Liangshan, China. *Remote Sens.* **2023**, *15*, 1513. [CrossRef]
53. Brenning, A. Spatial prediction models for landslide hazards: Review, comparison and evaluation. *Nat. Hazards Earth Syst. Sci.* **2005**, *5*, 853–862. [CrossRef]
54. Dou, J.; Yunus, A.P.; Dieu Tien, B.; Merghadi, A.; Sahana, M.; Zhu, Z.; Chen, C.-W.; Khosravi, K.; Yang, Y.; Binh Thai, P. Assessment of advanced random forest and decision tree algorithms for modeling rainfall-induced landslide susceptibility in the Izu-Oshima Volcanic Island, Japan. *Sci. Total Environ.* **2019**, *662*, 332–346. [CrossRef]
55. Wang, J.-F.; Zhang, T.-L.; Fu, B.-J. A measure of spatial stratified heterogeneity. *Ecol. Indic.* **2016**, *67*, 250–256. [CrossRef]
56. Lu, F.; Zhang, G.; Wang, T.; Ye, Y.; Zhen, J.; Tu, W. Analyzing spatial non-stationarity effects of driving factors on landslides: A multiscale geographically weighted regression approach based on slope units. *Bull. Eng. Geol. Environ.* **2024**, *83*, 394. [CrossRef]
57. Moran, P.A. Notes on continuous stochastic phenomena. *Biometrika* **1950**, *37*, 17–23. [CrossRef] [PubMed]
58. Polykretis, C.; Grillakis, M.G.; Argyriou, A.V.; Papadopoulos, N.; Alexakis, D.D. Integrating multivariate (GeoDetector) and bivariate (IV) statistics for hybrid landslide susceptibility modeling: A case of the vicinity of Pinios artificial lake, Ilia, Greece. *Land* **2021**, *10*, 973. [CrossRef]
59. Zhou, C.; Cheng, W.; Qian, J.; Bingyuan, L.I.; Zhang, B. Research on the Classification System of Digital Land Geomorphology of 1: 1,000,000 in China. *Geo-Inf. Sci.* **2009**, *11*, 707–724. [CrossRef]
60. Wang, N.; Cheng, W.; Wang, B.; Liu, Q.; Zhou, C. Geomorphological regionalization theory system and division methodology of China. *J. Geogr. Sci.* **2020**, *30*, 212–232. [CrossRef]
61. Cheng, W.; Zhou, C.; Chai, H.; Zhao, S.; Liu, H.; Zhou, Z. Research and compilation of the Geomorphologic Atlas of the People's Republic of China (1:1,000,000). *J. Geogr. Sci.* **2011**, *21*, 89–100. [CrossRef]
62. Yu, B.; Chen, W.; Feng, W.; Liu, K.; Ye, L. A case study of shallow landslides triggered by rainfall in Sanming, Fujian Province, China. *Environ. Earth Sci.* **2023**, *82*, 426. [CrossRef]
63. Liang, X.; Segoni, S.; Yin, K.; Du, J.; Chai, B.; Tofani, V.; Casagli, N. Characteristics of landslides and debris flows triggered by extreme rainfall in Daoshi Town during the 2019 Typhoon Lekima, Zhejiang Province, China. *Landslides* **2022**, *19*, 1735–1749. [CrossRef]
64. Li, Z.; Fotheringham, A.S.; Li, W.; Oshan, T. Fast Geographically Weighted Regression (FastGWR): A scalable algorithm to investigate spatial process heterogeneity in millions of observations. *Int. J. Geogr. Inf. Sci.* **2019**, *33*, 155–175. [CrossRef]
65. Sheng, Y.; Xu, G.; Jin, B.; Zhou, C.; Li, Y.; Chen, W. Data-Driven Landslide Spatial Prediction and Deformation Monitoring: A Case Study of Shiyan City, China. *Remote Sens.* **2023**, *15*, 5256. [CrossRef]

66. Zhou, X.; Wen, H.; Li, Z.; Zhang, H.; Zhang, W. An interpretable model for the susceptibility of rainfall-induced shallow landslides based on SHAP and XGBoost. *Geocarto Int.* **2022**, *37*, 13419–13450. [CrossRef]
67. 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]

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

A Fire Segmentation Method with Flame Detail Enhancement U-Net in Multispectral Remote Sensing Images Under Category Imbalance

Rui Zou [1], Zhihui Xin [1,*], Guisheng Liao [2], Penghui Huang [3], Rui Wang [1] and Yuhu Qiao [1]

[1] School of Physics and Electronic Information Technology, Yunnan Normal University, Kunming 650500, China; 2223090030@ynnu.edu.cn (R.Z.); 2324090020@ynnu.edu.cn (R.W.); 2324090013@ynnu.edu.cn (Y.Q.)
[2] National Laboratory of Radar Signal Processing, Xidian University, Xi'an 710071, China; liaogs@xidian.edu.cn
[3] School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China; huangpenghui@sjtu.edu.cn
* Correspondence: xinzhihui@ynnu.edu.cn

Abstract

Fire poses a serious threat to the global economy, environment, and social stability, highlighting the need for rapid and accurate fire detection. Remote sensing combined with deep learning has outperformed traditional fire assessment methods. However, in early fire stages, small flame areas, class imbalance, and weak feature extraction hinder detection accuracy. This study proposes an end-to-end segmentation model called Flame Detail Enhancement U-Net (FDE U-Net), using Landsat-8 multispectral remote sensing data. The model incorporates the self-Attention and Convolutional mixture (ACmix) module and the Convolutional Block Attention Module (CBAM) into the encoder of the Residual U-Net. ACmix integrates self-attention and convolution to capture global semantic features while maintaining computational efficiency, improving both contextual awareness and local detail. CBAM enhances flame recognition by weighting important channel features and focusing spatially on small flame areas, helping address the class imbalance problem. Additionally, Haar wavelet downsampling is applied to retain image detail and improve the detection of small-scale flame regions. Experimental results show that the FDE U-Net model exhibits robust performance in fire detection, accurately extracting flame regions even when their proportion is low and the background is complex. The F1 score reaches 95.97%, significantly improving the class imbalance problem.

Keywords: FDE U-Net; feature fusion; fire segmentation; multispectral remote sensing images

1. Introduction

Fires have recently become more frequent, destroying millions of hectares of forest annually, causing significant economic losses, deteriorating air quality, and reducing biodiversity. High temperatures and dry weather conditions caused by global warming increase the risk of fire [1]. Figure 1 shows the area burned by wildfires worldwide as of 12 June 2025. The figure clearly shows that fire is a major problem. Therefore, detecting forest fires rapidly and accurately is crucial. The key to fire detection is accurately identifying fire locations, a task strongly supported by computer vision technology [2].

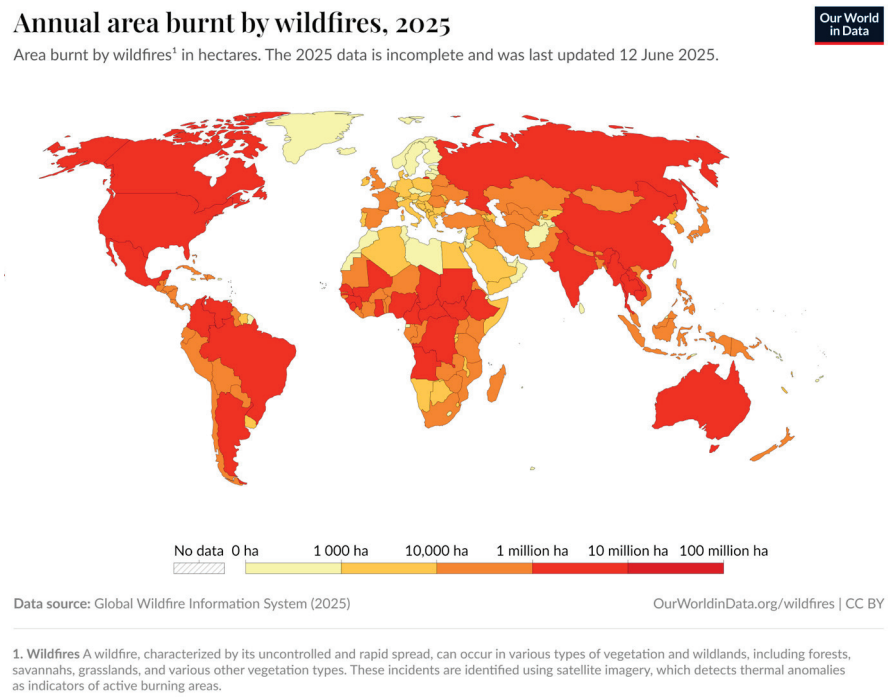


Figure 1. Area burnt by wildfires [3].

Traditional fire detection methods, such as using watchtowers located at high points to observe fire conditions, have improved efficiency and reduced labor intensity compared with manual inspections [4]. However, they are limited by the number of towers, terrain, and topography and have large blind spots, low accuracy, and high costs, making them difficult to use widely. Unmanned aerial vehicles can achieve large-scale, rapid, and effective real-time monitoring with a wide field of view and high efficiency. However, they have safety risks, cannot return images in real time, and are limited by night flights and environmental factors [5,6]. Whether it is UAV inspection or watchtower monitoring, vision-based algorithms have weak detection capabilities for small flame targets owing to camera pixel limitations, and these systems cannot achieve global coverage. Satellite-based systems utilize satellites to provide services. Compared to terrestrial systems, they can achieve global coverage and provide secure connections that are not affected by physical or weather obstacles, making them a viable option for forest fire detection [7,8].

Diverse types of satellite imagery are widely used in fire detection as the basis for monitoring and analysis. The moderate resolution imaging spectroradiometer (MODIS) and the visible infrared imaging radiometer suite (VIIRS) sensors are both capable of capturing multiple spectral bands, enabling effective identification of fire-related thermal radiation and smoke characteristics [9,10]. Cardíl et al. conducted the first systematic analysis of the rate of spread (ROS) of large wildfires in Northwestern Europe using VIIRS satellite active fire data, revealing the influence of different land cover types and seasons on fire spread and providing critical insights for risk management and response strategies in this emerging fire-prone region [11]. Zhang et al. jointly used VIIRS 375 m I-Band and 750 m M-Band data to detect and evaluate fires and fire radiated power (FRP) and proposed a VIIRS-IM method that improved the sensitivity and accuracy of small-area fire detection, which provided more fire pixels and more accurate FRP records compared to MODIS data [12]. Hong et al. proposed the FireCNN model based on Himawari-8 satellite images using multiscale convolution and residual concatenation to efficiently extract the accurate features of fire points with an accuracy rate 35.2% higher than that of the traditional threshold method [13]. These satellite data are used for active fire detection

(AFD). However, their low spatial resolution can reduce AFD accuracy. The Landsat-8 satellite, with its high spatial and temporal resolution, overcomes this problem [14].

In addition to relying on reliable image data, the exploration of accurate and efficient fire detection methods has become a key research topic. Researchers have proposed and applied various popular methods, most of which focus on color and shape features [15,16]. For example, Çelik et al. used the YCbCr color space to better distinguish brightness and chromaticity and constructed a universal flame pixel classification chromaticity model, which effectively reduced illumination change impact with high detection and low false alarm rates [17]. Chen et al. proposed a fire multi-feature fusion recognition algorithm, which outperforms the traditional method in performance, but the recall is slightly lower than the deep learning method [18]. Jang et al. combined thresholding, random forest, and postprocessing techniques to enhance small fire detection accuracy [19]. Although the above methods have achieved good results, they suffer from poor generalization ability and suboptimal performance in fire detection in complex backgrounds, making them insufficient for most fire detection tasks.

Owing to its powerful feature extraction capabilities and the ability to process high-dimensional data, deep learning has achieved remarkable results in areas such as image classification [20], target detection [21,22], and image segmentation [23,24]. The use of deep learning methods for fire detection is considerably effective [25,26]. Sun et al. significantly reduced the computational complexity of YOLO v7 while maintaining high fire detection accuracy by improving activation functions, lightweight modules, and backbone network reconstruction [27]. Hu et al. proposed a network architecture called AF-Net, which significantly improves the detection accuracy of imbalanced fire data in UAV remote sensing images by optimizing the sampling strategy and enhancing the object-contextual representation (OCR) structure. Numerous studies have focused on deep-learning-based approaches and used Landsat-8 data to detect fires [28]. Seydi et al. proposed Fire-Net, a forest fire detection approach based on deep learning, which utilizes residual and separable convolutional modules. The model was trained on 578 Landsat-8 images and tested on multiple scenes, achieving high accuracy in all cases [29]. Han et al. constructed the BA-BS dataset based on Landsat multispectral remote sensing imagery and auxiliary environmental data and proposed a transformer-based change detection model incorporating auxiliary information, which demonstrated excellent performance in burned area identification and burn severity classification tasks [30]. The U-Net model has become an important benchmark model in image segmentation owing to its symmetrical structure, fully convolutional network, multiscale feature fusion, and end-to-end training. It can capture finer details and produce more accurate results [31–33]. Xu et al. proposed the MSU-Net model, which introduced self-attention and multi-scale dilated convolutional structures based on U-Net for high-precision extraction of tea-oil plantation areas in high-resolution UAV images. This method places special emphasis on global modeling and small target recognition capabilities, demonstrating robustness and practicality under hilly and complex terrain conditions [34]. Wang et al. proposed Smoke-Unet by improving the U-Net method and combining spatial-based and channel-based attention with residual blocks. A smoke dataset based on multispectral remote sensing was developed, encompassing multiple years, diverse seasons, geographic regions, and land cover categories. The accuracy rate reached 92.3%, which was better than the basic methods such as U-Net, FCN [35], PSPNet [36], etc [37]. Afsar et al. proposed an encoder–decoder-encoder–decoder structure, ResWnet, which reduced feature loss during sampling and retained more detailed features. The accuracy of ResWnet was 95%, and the F1 score was 94.2% [38]. Compared with traditional methods, the above models achieved higher detection accuracy. However, limited feature expression ability, insufficient capture of microscopic targets, and inadequate processing of

local features remain problems. One of the main reasons for the lack of previous research on the use of satellite imagery for fire detection is the lack of suitable datasets. de Almeida Pereira et al., referring to the method proposed in [39–41], introduced a large-scale multispectral dataset, large-scale active fire detection (LAFD), that can be used for fire detection in 2021 [42]. It contained more than 150,000 image blocks, providing strong support for fire detection.

Owing to the 30 m resolution, wide coverage, and multiple bands of Landsat 8 multispectral data, substantial storage and computational resources are required, which limits the processing speed. Additionally, the images are often covered by clouds and cloud shadows, which affect the observations. The fire area is usually mixed with a complex background. The existing model has limited feature expression ability, and it is difficult to accurately segment the fire area. The shapes and sizes of flames vary, and existing algorithms have limited capabilities in capturing small targets and adequately processing local features. Therefore, this study applies semantic segmentation to fire detection based on an LAFD dataset [42] provided by de Almeida Pereira et al. and proposes a new segmentation network, Flame Detail Enhancement U-Net (FDE U-Net). It achieves the adaptive selection and mixing of input features, effectively capturing subtle features in the early stages of a fire. When analyzing fire images, local information is usually more important than global information, as it can provide more accurate pixel-level information regarding fire areas. The main contributions of this study are as follows:

(1) FDE U-Net comprehensively uses advanced feature extraction and enhancement techniques to increase the depth and breadth of feature extraction, optimize information transfer, and significantly improve detection accuracy and efficiency. Multilevel feature processing and fusion enable the model to more accurately identify and locate flame details when it faces complex multispectral data.

(2) The self-Attention and Convolutional mixture (ACmix) and Convolutional Block Attention Module (CBAM) are integrated into the encoder part of ResU-Net to achieve adaptive selection and mixing of the input features, which significantly improves the category imbalance problem and, thus, the detection accuracy of the model for the flame region.

(3) The Haar wavelet downsampling (HWD) can retain more boundary, texture, and detail information while reducing the size of the feature map using the Haar wavelet transform. The feature representation and detail capture capabilities are enhanced, effectively capturing tiny features in the early stages of a fire.

The following sections further introduce and discuss the proposed network. Section 2 provides a detailed introduction to the basic structure of the proposed network. The details of the dataset are presented in Section 3. Section 4 presents the experimental results and discusses them from both quantitative and qualitative perspectives. Section 5, as the final part, draws the conclusion of this paper.

2. Model

2.1. Network Architecture

We optimize the U-Net architecture and propose a dual-path encoder–decoder-based method, FDE U-Net, to improve its feature learning capability and the efficiency of contextual information acquisition for the accurate segmentation of fire regions. Figure 2 shows the overall network architecture of the FDE U-Net.

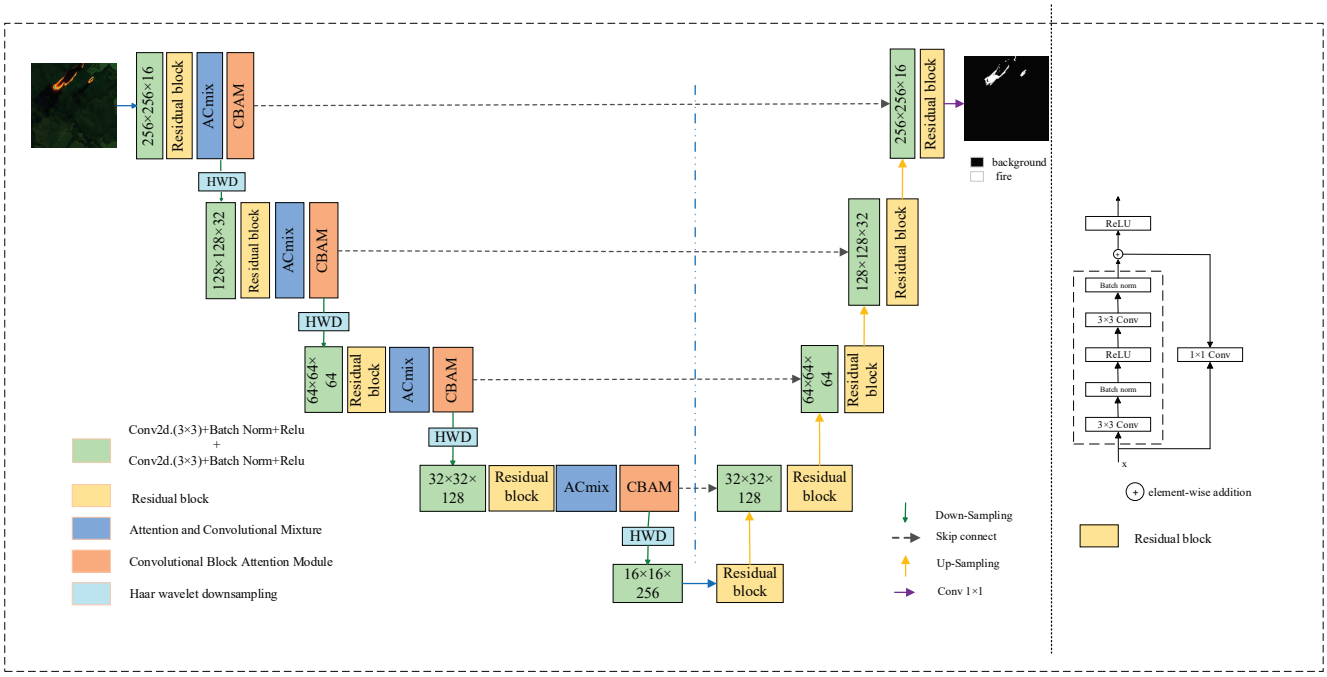


Figure 2. Overall architecture of FDE U-Net.

FDE U-Net acts as an end-to-end encoder–decoder network, with the encoder portion receiving the input image and the number of channels extended to 16, enabling the network to extract richer feature information from the input data, which helps to capture spatial and spectral features in multi-band images. The encoder uses four residual blocks, which effectively mitigates the gradient-vanishing problem in deep network training, improving the model's ability to extract deep features. The ACmix and CBAM are integrated into each downsampling layer. The ACmix module adaptively blends cross-channel features between multispectral bands by combining convolution and self-attention operations to promote effective interaction and fusion of features between different channels. This enables the model to effectively integrate the differentiated spectral characteristics between different bands, enriches the feature expression, and improves the model's ability to understand complex image content. To further improve the model's performance, CBAM refines the weighting of feature maps in two dimensions, space and channel, such that the model can focus on key areas in the image. CBAM can adjust its focus based on the response characteristics of different bands. For example, in the thermal infrared band, it enhances the fire area, reduces background noise, and effectively separates the fire. In the spatial dimension, it focuses on the fire's morphology and boundaries, thereby improving segmentation accuracy. Regarding the downsampling strategy, the traditional maximum pooling operation is abandoned, and HWD is adopted. The HWD utilizes the multiscale analysis capability of the wavelet transform and reduces the resolution of feature maps through the Haar wavelet transform. While reducing the resolution, the fire edges, texture details, and cross-band contextual association information in the multispectral image are retained, which retains more image detail information and provides richer contextual information for subsequent segmentation tasks.

Specifically, the decoder gradually restores the image size through upsampling. After each up-sampling, the spatial resolution of the feature map increased, whereas the number of feature channels is halved, yielding deep features that are skip-connected to the corresponding output features from the encoder. By leveraging skip connections, the model can integrate low-level details with high-level semantics, which supports accurate boundary reconstruction and mitigates the degradation of crucial features when restoring

spatial resolution. In the final layer, classification is performed using a 1×1 convolution followed by a Softmax activation function, resulting in fine segmentation. Simultaneously, batch normalization and dropout layers are added after each convolution layer to avoid overfitting. Finally, a binary cross-entropy (BCE) loss function is used to measure the difference between the model's predicted probability and the actual category to evaluate the model's performance. In binary segmentation tasks, the binary cross-entropy (BCE) loss is formulated as follows:

$$BCE_{loss} = -\frac{1}{n}\sum_{i=1}^{n}(y_i log(p) + (1 - y_i) log(1 - p)) \quad (1)$$

where p is the probability of the fire class, $1 - p$ is the probability of the non-fire class, and y_i is the pixel label.

2.2. Self-Attention and Convolutional Mixture (ACmix)

Convolution and self-attention are two powerful representation-learning techniques. For a standard convolution operation, Figure 3a shows a 3×3 convolution decomposed into nine 1×1 convolutions, followed by shift summation [43]. First, a linear projection of the input feature map is performed using the kernel weights at a specific location; the projected feature map is then shifted and aggregated based on the position. For the self-attention operation, as shown in Figure 3b, the input features are first projected into queries, keys, and values using a 1×1 convolution, after which the attention weights are computed and the values are aggregated. As shown in the figure, the first step of both operations includes a 1×1 convolution. ACmix combines the advantages of convolution and self-attention to achieve the dual functions of the deep extraction of spatial features and efficient aggregation of global information.

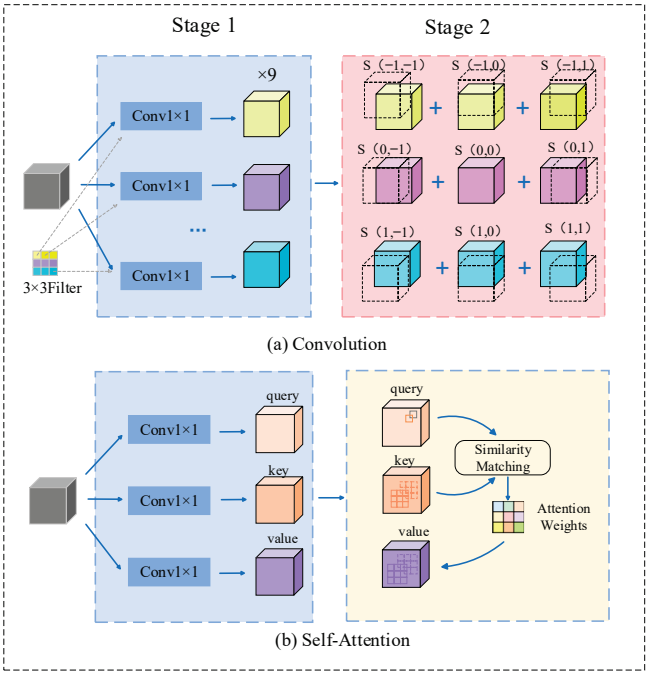


Figure 3. (**a**) A 3×3 convolution decomposition and shift summation process; (**b**) self-attention operation process.

The second step is to reuse and aggregate the intermediate features in different modes (self-attention and convolution). When obtaining the self-attention feature, the intermediate feature set is divided into N groups. Each group contains three feature maps, and the three feature maps are used as *query*, *key*, and *value* matrices to obtain the self-attention feature

$f_{self-attention}$. In obtaining the convolution features, assuming that the convolution kernel size is k, a fully connected layer is used to obtain k^2 feature maps, and the convolution feature f_{conv} is calculated and summed. To eliminate the inefficiency of tensor shift operations, we employ depthwise convolutions with fixed kernels as a lightweight and effective alternative. If we denote the convolution kernel (kernel size k = 3) as

$$K_c = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \forall c \tag{2}$$

The corresponding output can be formulated as

$$\begin{aligned} f_{c,i,j}^{(dwc)} &= \sum_{p,q \in \{0,1,2\}} K_{c,p,q} f_{c,i+p-\lfloor k/2 \rfloor, j-q-\lfloor k/2 \rfloor} \\ &= f_{c,i-1,j-1} = \widetilde{f}_{c,i,j}, \forall c, i, j \end{aligned} \tag{3}$$

where c represents each channel of the input feature. $\widetilde{f}_{c,i,j}$ represents the calculation result of tensor shifts.

Therefore, with carefully designed kernel weights for specific shift directions, the convolution outputs are equivalent to the simple tensor shifts. This will improve the actual efficiency of the module during inference. Finally, ACmix reuses intermediate features to avoid repeated and complex double-projection operations on the same feature set, thereby reducing redundant calculations, improving computational efficiency, and reducing model complexity while maintaining high performance. Finally, using two learnable scalars, the self-attention and convolution features are organically combined to obtain the final fusion feature. Figure 4 shows the specific operation process of ACmix. The expression is as follows:

$$f_{out} = \alpha \times f_{self-attention} + \beta \times f_{conv} \tag{4}$$

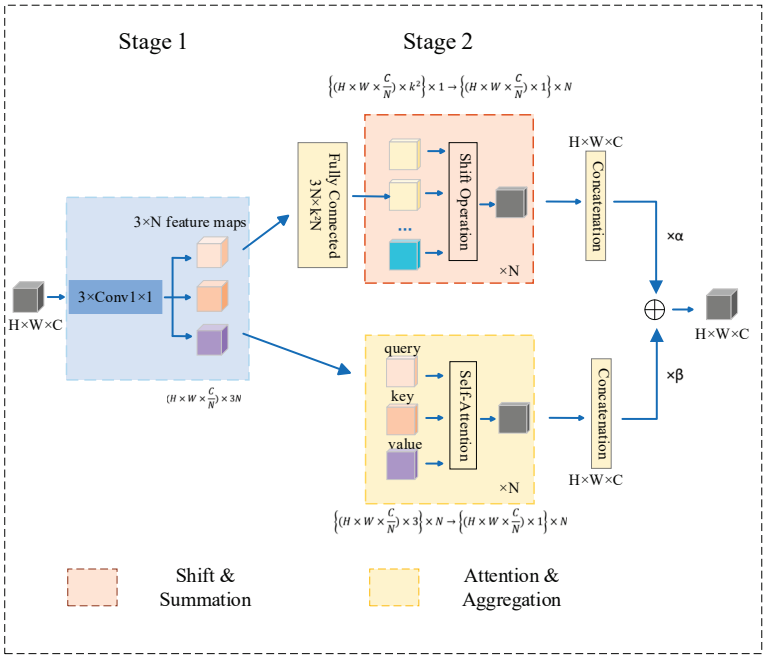


Figure 4. Overall structure of ACmix.

2.3. Convolutional Block Attention Module (CBAM)

During the low-level feature fusion between the encoder and decoder, the CBAM attention mechanism is employed to enhance the extraction of valuable details, guiding

the model to better focus on flame regions and boosting its overall performance. CBAM combines spatial and channel attention in a convolutional block [44]. Figure 5 shows its structure. After being processed by the convolutional layer, the output is first passed through the channel attention mechanism to apply weighting, and then refined by the spatial attention mechanism to produce the final output. The CBAM is introduced to enhance the features associated with the flame area. Background regions unrelated to fire are effectively suppressed in the attention feature map, unlike in the raw features extracted by the base network.

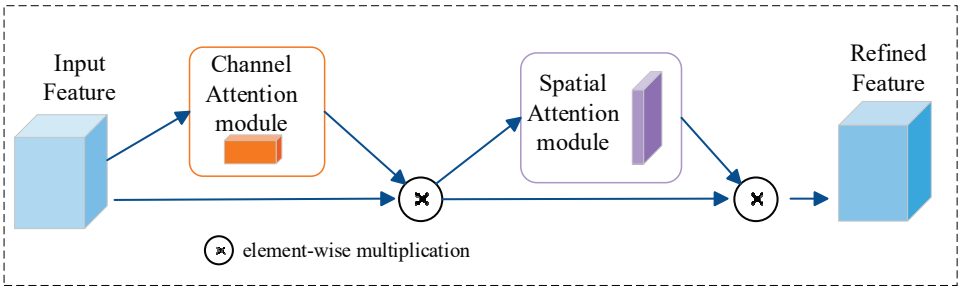


Figure 5. Convolutional block attention module.

Figure 6 shows the details of the channel attention implementation. Input feature map F is compressed in spatial dimension by global average pooling (GAP) and global maximum pooling (GMP) operations to obtain one-dimensional vectors F^c_{avg} and F^c_{max}, corresponding to the feature representations obtained through average and maximum pooling, respectively. The output sizes of the GAP and GMP are $1 \times 1 \times C$. They are then fed into a shared multilayer perceptron (MLP), comprising two fully connected layers, producing outputs denoted as M_{avg} and M_{max}. The outputs of the GAP and GMP after the MLP are summed element-by-element to obtain the channel attention weight map M_c. The original input feature map F is multiplied by the channel attention weight map M_c to adjust the weight of each channel. The channel attention mechanism can selectively enhance the important feature channels and suppress irrelevant or redundant channels by calculating the importance weight of each channel. The computation of channel attention is defined as follows:

$$M_{avg} = W_1(W_0(F^c_{avg})) \tag{5}$$

$$M_{max} = W_1(W_0(F^c_{max})) \tag{6}$$

$$M_c = \sigma(M_{avg} + M_{max}) \tag{7}$$

where σ represents the sigmoid function, and W_0 and W_1 are the two input-shared MLP weight matrices.

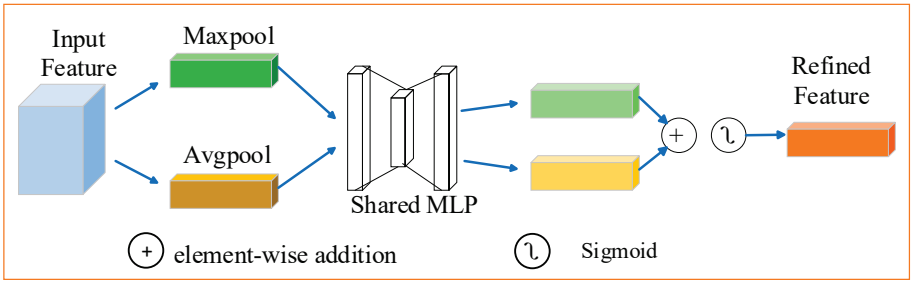


Figure 6. Channel attention module.

The implementation details of spatial attention are shown in Figure 7. It mainly calculates the importance weight of each position to perform weighted adjustment on the input feature map and highlight the key spatial features. The feature map produced by the channel attention mechanism is first utilized as the input feature representation for this module. Global average pooling (GAP) and global maximum pooling (GMP) are performed on the input feature map, and pooling is performed along the channel dimension, respectively, to obtain two single-channel feature maps F^S_{avg} and F^S_{max}. The output sizes of GAP and GMP are both H × W × 1. Then the single-channel feature maps generated by GAP and GMP are concatenated in the channel dimension to obtain a feature map of size H × W × 2. Finally, a 7 × 7 convolutional layer is applied to the concatenated feature map to produce the spatial attention weight map M_S. The original input feature map F is multiplied by the spatial attention weight map M_S to perform a weighted adjustment on each position. Spatial attention is calculated as follows:

$$\begin{aligned} M_S &= \sigma\left(f^{7\times7}([AvgPool(F); MaxPool(F)])\right) \\ &= \sigma\left(f^{7\times7}\left(\left[F^S_{avg}; F^S_{max}\right]\right)\right) \end{aligned} \tag{8}$$

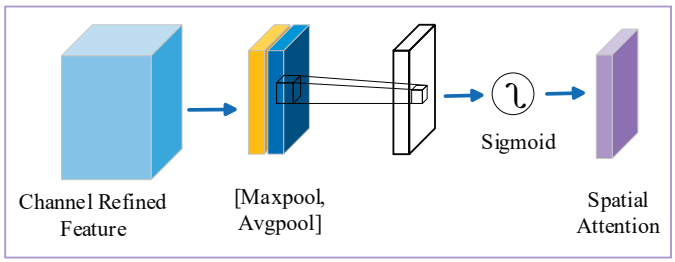


Figure 7. Spatial attention module.

2.4. Haar Wavelet Downsampling (HWD)

Downsampling techniques like max pooling are commonly employed in convolutional neural networks (CNNs) to consolidate local feature information, enlarge the receptive field, and reduce computational complexity. However, in semantic segmentation tasks, performing pooling operations over local neighborhoods can inevitably lead to the loss of crucial spatial information, thereby weakening the model's ability to preserve details. HWD retains more detail and edge features by decomposing the image. Compared to traditional maximum pooling or average pooling, HWD can better preserve the original information of the image and reduce information loss [45].

The HWD consists of two main parts (Figure 8). The first part of the lossless feature-encoding block reduces the spatial resolution of the feature map using the Haar wavelet transform. The input image is divided into four sub-bands using Haar wavelet decomposition: a single low-frequency component, denoted as LL, along with three high-frequency components, namely LH, HL, and HH. By downsampling the above components, the feature map size is reduced to half of the original size, whereas the number of feature map channels is increased fourfold. The size of the output feature map was $\frac{H}{2} \times \frac{W}{2} \times 4C$. In the second stage of the feature representation learning block, redundant information is suppressed using a standard 1 × 1 convolution, followed by batch normalization and ReLU activation. Standard convolution is applied to modify the channel dimensions of the feature map, aligning it with subsequent layers and enabling them to more effectively capture meaningful and representative features.

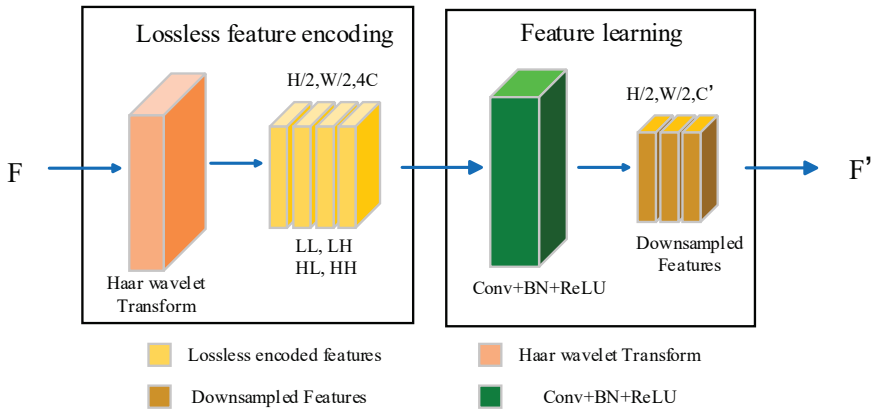


Figure 8. Haar wavelet downsampling.

3. Datasets and Evaluation Metrics

3.1. Datasets

Pereira et al. released a new large multispectral dataset called LAFD in 2021 [40]. This dataset was extracted from the Landsat-8 satellite launched by NASA and included wildfires that occurred at various locations worldwide between August and September 2020. The samples covered various areas, including deserts, rainforests, farmlands, water, snow, clouds, cities, and mountains. All image blocks in the dataset were 256 × 256 pixels. These image blocks were 10-band, 16-bit TIFF images. The 10 bands included the coastal band (COASTAL/AEROSOL), three RGB bands, near-infrared (NIR), shortwave infrared (SWIR 1 and SWIR 2), cirrus band (Cirrus), and thermal infrared band (TIR1 and TIR2). The channels were b1–b11, excluding b8 (panchromatic channel), with a spatial resolution of 30 m.

A significant category imbalance is observed between the fire and non-fire pixels in the dataset. We select image patches with a high and low proportion of fire pixels to detect small fire regions. A total of 1100 image blocks are selected from the huge LAFD dataset, and these data are horizontally and vertically flipped, ultimately generating 3300 image blocks to enhance the diversity of the dataset. The dataset was divided with 80% used for the training and validation processes and 20% reserved for testing to ensure the independence and effectiveness of the model evaluation. Our dataset refers to the three algorithms discussed in [39–41]. We use a "three-round, two-win" voting scheme (if at least two sets of conditions agree that it was a fire pixel, it is considered a fire) to generate the dataset. In Figure 9, (a) shows one of the fire scenes; (b), (c), and (d) represent the masks generated by the three algorithms; and (e) represents the voting result. The three algorithms are as follows:



(a) **(b)** **(c)** **(d)** **(e)**

Figure 9. Examples of masks for each algorithm. (**a**) Original image. (**b**) Schroeder. (**c**) Murphy. (**d**) Kumar and Roy. (**e**) Voting.

(1) Schroeder et al. [39]

Seven Landsat-8 channels (c1 to c7) are used to identify the fire pixels. Unambiguous fire pixels must meet the following criteria:

$$\begin{array}{l}[(R_{75} > 2.5)and(\rho_7 - \rho_5 > 0.3)and(\rho_7 > 0.5)]or\\ [(\rho_6 > 0.8)and(\rho_1 < 0.2)and(\rho_5 > 0.4or\rho_7 < 0.1)]\end{array} \quad (9)$$

For each fire candidate pixel, the neighborhood statistics of a 61 × 61 pixel window are adjusted as follows:

$$\begin{array}{l}(R_{75} > 1.8)and(\rho_7 - \rho_5 > 0.17)and(R_{76} > 1.6)and\left(R_{75} > \mu_{R_{75}} + max\left(3\sigma_{R_{75}}, 0.8\right)\right)\\ and\left(\rho_7 > \mu_{\rho_7} + max\left(3\sigma_{\rho_7}, 0.08\right)\right)\end{array} \quad (10)$$

(2) Murphy et al. [41]

A set of context-independent conditions is proposed so that it does not rely on the statistical data calculated from each pixel's neighborhood. Only channels c5, c6, and c7 of Landsat-8 are used. The unambiguous fire pixels must satisfy the following conditions:

$$(R_{76} \geq 1.4)and(R_{75} \geq 1.4)and(\rho_7 \geq 0.15) \quad (11)$$

Potential fires located within the vicinity of an identified fire (within a 3 × 3 window) are classified as fires. A potential fire meets the following conditions:

$$(R_{65} \geq 2)and(\rho_6 \geq 0.5)or((\rho_7\ is\ saturated)or(\rho_6\ is\ saturated)) \quad (12)$$

(3) Kumar and Roy [40]

Using channels c2 to c7, unambiguous fire pixels must meet the following conditions:

$$\rho_4 \leq 0.53\rho_7 - 0.214 \quad (13)$$

Potential fire pixels are identified based on the following conditions:

$$(\rho_4 \leq 0.53\rho_7 - 0.125)or(\rho_6 \leq 1.08\rho_7 - 0.048) \quad (14)$$

$$\left(R_{75} > \mu_{R_{75}} + max\left(3\sigma_{R_{75}}, 0.8\right)\right)and\left(\rho_7 > \mu_{\rho_7} + max\left(3\sigma_{\rho_7}, 0.08\right)\right) \quad (15)$$

where ρ_k represents the reflectivity of channel k. R_{ij} is the ratio of the reflectance of channels i and j. μ and σ represent the mean and standard deviation of the pixels in the 61 × 61 window pixel neighborhood.

3.2. Evaluation Metrics

We evaluate the architecture based on the main metrics used in semantic segmentation: precision (P), recall (R), F1 score (F1), and intersection over union (IoU). Precision refers to the proportion of pixels that belong to the fire region among all the pixels predicted by the model as fire regions. Recall measures the proportion of correctly identified fire pixels relative to the total number of actual fire pixels. The F1 score, defined as the harmonic mean of precision and recall, reflects the balance between these two metrics, with a higher value indicating better overall performance. However, owing to the severe imbalance between fire and non-fire categories, a more convincing criterion is required to understand the actual performance of the model. Therefore, we use the IoU, which measures the degree of overlap between the predicted and real segmentation results. The IoU and F1 score are complementary in evaluating model performance. The IoU focuses on the

spatial overlap between the predicted and ground truth results, whereas the F1 score emphasizes the accuracy and completeness of the detection results. Therefore, we select the IoU and F1 score as the main metrics to evaluate the segmentation performance, as they comprehensively reflect the model's performance from different perspectives. The metrics are defined as follows:

$$Precision = \frac{TP}{TP + FP} \tag{16}$$

$$Recall = \frac{TP}{TP + FN} \tag{17}$$

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{18}$$

$$IoU = \frac{TP}{TP + FN + FP} \tag{19}$$

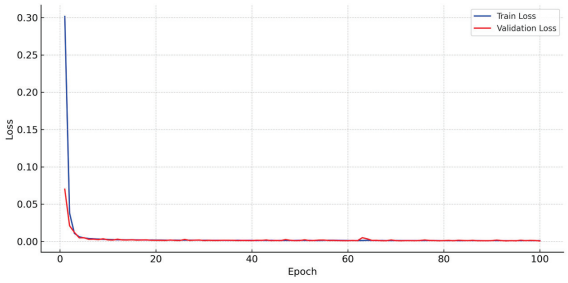
where TP, FP, and FN refer to the number of samples correctly predicted to be in the positive category (correct detection), the number of negative samples incorrectly predicted to be in the positive category (false positive), and the misclassified positive samples categorized as negatives, referred to as false negatives, respectively.

4. Analysis of Fire Detection Results

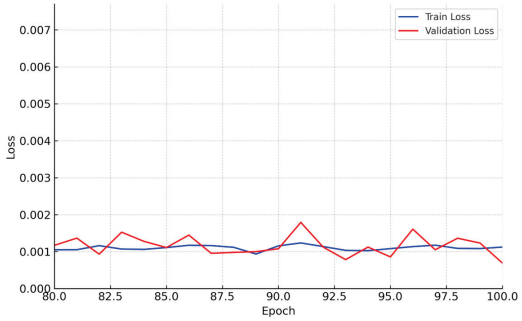
4.1. Training Parameter Settings

All experiments were implemented using PyTorch 1.12.1 and conducted on an NVIDIA GeForce RTX 3080 Ti 12 GB GPU. The training process employed the Adam optimizer, initialized with a learning rate of 1×10^{-3}, and BCE was used as the loss function to train each architecture for 100 epochs.

Figure 10 shows the loss curve of the proposed architecture. It can be seen that the loss curve decreases and eventually stabilizes. Early stopping was applied to prevent overfitting, and the training was stopped when the validation loss did not improve within five epochs. The batch size used during training was set to 32 to balance the memory capacity and training speed.



(**a**) Original image



(**b**) Enlarged image of epoch 80–100

Figure 10. FDE U-Net loss curve.

4.2. Ablation Experiments

We conduct a detailed performance analysis through ablation experiments to verify the effectiveness of the adopted modules. Table 1 presents the ablation experimental results based on ResU-Net. Before each downsampling of the ResU-Net encoder, ACmix, HWD, and CBAM are added separately, and two of them are added at the same time, which are denoted as Improvements 1 to 6, respectively. After adding ACmix and CBAM, the maximum pooling is replaced by HWD and expressed as FDE U-Net. As shown in Table 1, significant improvements are observed in four performance indicators: precision (P), recall (R), IoU, and F1 score. ACmix + CBAM + HWD provides the greatest improvement, with all metrics reaching an optimal level.

Table 1. Ablation experiment results based on ResU-Net.

Method	ACmix	CBAM	HWD	P (%)	R (%)	IoU (%)	F1 (%)
ResU-Net				96.03	84.22	81.31	89.69
Improvement 1	√			95.28	95.28	90.99	95.28
Improvement 2		√		96.58	93.59	90.60	95.06
Improvement 3			√	96.49	93.22	90.16	94.82
Improvement 4	√	√		96.22	94.83	91.43	95.53
Improvement 5	√		√	96.05	95.66	92.04	95.86
Improvement 6		√	√	96.43	94.06	90.89	95.23
FDE U-Net	√	√	√	**97.05**	**95.87**	**92.25**	**95.97**

Specifically, ACmix enhances the model's fusion of cross-channel features between multispectral bands by adaptively selecting and mixing input features. Compared to ResU-Net, Improvement 1 achieves an 11.06% increase in recall, enhancing the model's sensitivity to complex backgrounds and diverse fire characteristics. However, its precision decreases by 0.75%, likely due to the model amplifying non-fire disturbances with similar spectral features. CBAM dynamically adjusts the weights of feature maps by integrating channel and spatial information. Channel attention assigns higher weights to important spectral bands, while spatial attention highlights critical regions, enabling the model to enhance fire-related feature regions and feature channels in fire detection tasks, and suppressing the interference of extraneous information. Improvement 2 improves precision by 0.55% compared to ResU-Net. HWD effectively mitigates feature blurring caused by traditional pooling operations, preserving finer details and boundary information during the downsampling process. Meanwhile, Improvement 3 increases recall by 9% while maintaining high precision, surpassing ResU-Net. Furthermore, the progressive integration of the three modules consistently enhances IoU and F1 scores demonstrating the effectiveness of the modular optimization strategy. This modular optimization process progressively improves model accuracy, strengthens its ability to handle complex scenes and fine-grained segmentation, and ensures a better balance between precision and recall. The specific data show that, compared with the baseline model ResU-Net, precision increases by 1.02%, recall increases by 11.65%%, IoU increases by 10.94%%, and F1 score increases by 6.28%, respectively. The gradual addition of these modules positively and significantly contributes to the overall performance of the model.

Figure 11 shows the segmentation results of the different models for the fire area. We mark false positives in red and false negatives in blue. As shown in Figure 11c–f, all the methods detect the correct location of the fire, with the main errors occurring at the edges of the fire and non-fire areas contained within the large fire area. Figure 11c shows the ResU-Net baseline detection results, with false positives at the edges and non-fire areas between fires misclassified as fire. Figure 11d shows that recall increases and false negatives

decrease with ACmix, indicating its role in feature extraction. Adding only CBAM reduces both false negatives and false positives, highlighting its effectiveness in reinforcing critical information. The combination of ACmix and CBAM balances precision and recall in feature extraction. Haar wavelet downsampling improves edge segmentation. Our method reduces false positives and false negatives, as shown in the elliptical box in Figure 11f, where the non-fire area between fire regions is correctly segmented with minimal false positives. Experimental results further indicate that removing any individual module adversely affects segmentation performance, whereas the combined implementation of these modules significantly enhances segmentation effectiveness.

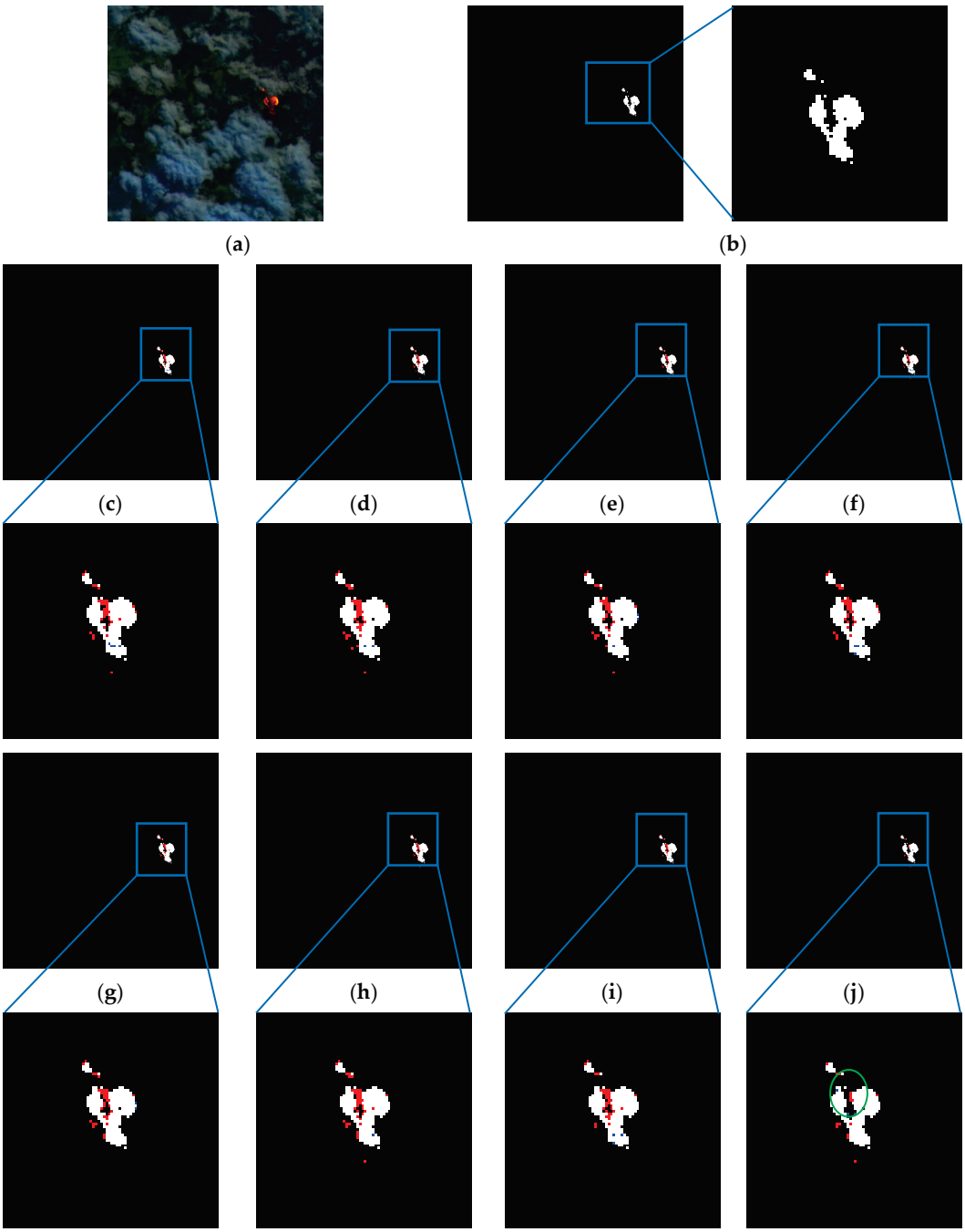


Figure 11. (**a**) Original image. (**b**) Ground truth. (**c**) ResU-Net. (**d**) Improvement 1. (**e**) Improvement 2. (**f**) Improvement 3. (**g**) Improvement 4. (**h**) Improvement 5. (**i**) Improvement 6. (**j**) FDE U-Net.

4.3. Fire Detection Results

4.3.1. Overall Fire Detection Results Analysis

In the ablation experiment described in Section 4.2, the effectiveness of FDE U-Net is demonstrated. To comprehensively evaluate its performance, we compare the proposed method with several state-of-the-art segmentation networks, including U-Net, ResU-Net, U-Net3+, SwinU-Net, and WET-UNet. Table 2 presents the fire detection performances of the different models. The proposed method has the best performance among all models, with IoU and F1 scores of 92.25% and 95.97%, respectively. Compared with the suboptimal SwinU-Net, the IoU increases by 0.39% and the F1 score improves by 0.21%. U-Net, as the basic segmentation network, is lower than the improved model in all metrics, especially in the IoU and F1 metrics, with room for improvement. ResU-Net exhibits improved recall but slightly decreased precision. ResU-Net enhances feature transfer through residual connections, improving model stability, convergence speed, and feature reuse. However, this approach can excessively smooth input features, leading to a reduction in recall. Moreover, the spectral bands of Landsat 8 exhibit varying sensitivities to fire, posing a challenge, as ResU-Net diminishes the influence of key bands such as SWIR-2, thereby weakening the detection of small or subtle fire-related features. U-Net3+ enhances the performance through more skip connections and multiscale feature fusion, with a precision of 96.49%. However, the detection of small objects or edges may not be sufficiently sensitive, resulting in low recall rates. SwinU-Net leverages global information to capture the overall fire extent, enhancing recall over a larger spatial range and detecting more fire areas. However, it may mistakenly classify pixels adjacent to fire zones as fire, leading to a decrease in precision compared to U-Net3+. Additionally, this approach involves a large number of parameters, which can increase computational complexity. By incorporating wavelet transform, WET-UNetcompresses information efficiently and enhances feature extraction, improving edge detection and reducing redundancy, which leads to better precision.

Table 2. Performance comparison of FDE U-Net and other CNN methods.

Method	P (%)	R (%)	IoU (%)	F1 (%)	Para (M)	FLOPS	Inference Time	FPS
U-Net [31]	94.91	84.43	80.71	89.32	1.95	5.44 G	17.33 ms	57.69
ResU-Net [46]	96.03	84.22	81.31	89.69	2.03	4.23 G	19.81ms	50.45
U-Net3+ [47]	96.49	90.16	87.30	93.22	2.31	2.02 G	18.05 ms	55.39
SwinU-Net [48]	95.85	95.66	91.86	95.76	8.73	4.42 G	24.30 ms	41.15
WET-UNet [49]	96.99	93.38	90.76	95.15	6.32	3.16 G	26.10 ms	38.31
FDE U-Net	**97.05**	**95.87**	**92.25**	**95.97**	2.25	4.16 G	20.13 ms	49.7

FDE U-Net offers stronger feature expression and processing. The short-wave infrared band of Landsat 8 is sensitive to fires, and Haar wavelet decomposition preserves high-temperature features during downsampling, enhancing the extraction of smoke edges and fire contours. The synergy between ACmix and CBAM captures the global dependency between multispectral bands, highlights the key bands through channel weight adjustment, and focuses the fire diffusion direction in the spatial dimension. The integration of global and local information enables the model to cope with complex segmentation tasks more effectively, and improves the segmentation precision and detail restoration ability in the case of category imbalance. The parameter amount of FDE U-Net is 2.25M. Compared to SwinU-Net and WET-UNet, the number of parameters is significantly reduced, which gives the proposed method a higher practical value in scenarios with limited computing resources. We further compare inference time, FLOPs (Floating-point Operations), and FPS (frames per second) of different models. FDE U-Net, based on the comprehensive

enhancement of multiple modules, still effectively controls the model's computational workload and inference time. FLOPs are reduced by about 23.53% compared to traditional U-Net, significantly reducing the computational burden of the model. Under the premise of integrating multiple advanced modules, FDE U-Net's inference time is still controlled at around 20 ms, which is faster than SwinU-Net (24.30 ms) and WET-UNet (26.10 ms), indicating that its structural design has good inference efficiency while being compatible with high expressiveness.

Regarding the fire detection task, the main challenges are twofold: (1) the fire area accounts for a small proportion compared to the background, and there is a serious category imbalance problem. (2) Fire regions often have irregular shapes and boundaries, which pose significant challenges for accurate segmentation. We select three representative scenarios, images 1 to 4, which contain a centralized target distribution scenario and a decentralized distribution scenario, as shown in Figure 12. Figures 13–16 show the fire segmentation results of images 1 to 4 using several advanced models. For ease of observation, false positives are marked in red and false negatives are marked in blue. Table 3 presents the statistics for the correctly and incorrectly detected pixels.

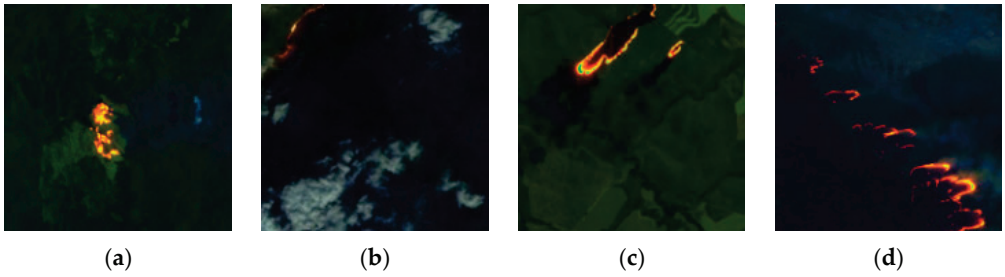


Figure 12. (**a**) Image 1, (**b**) image 2, (**c**) image 3, (**d**) image 4.

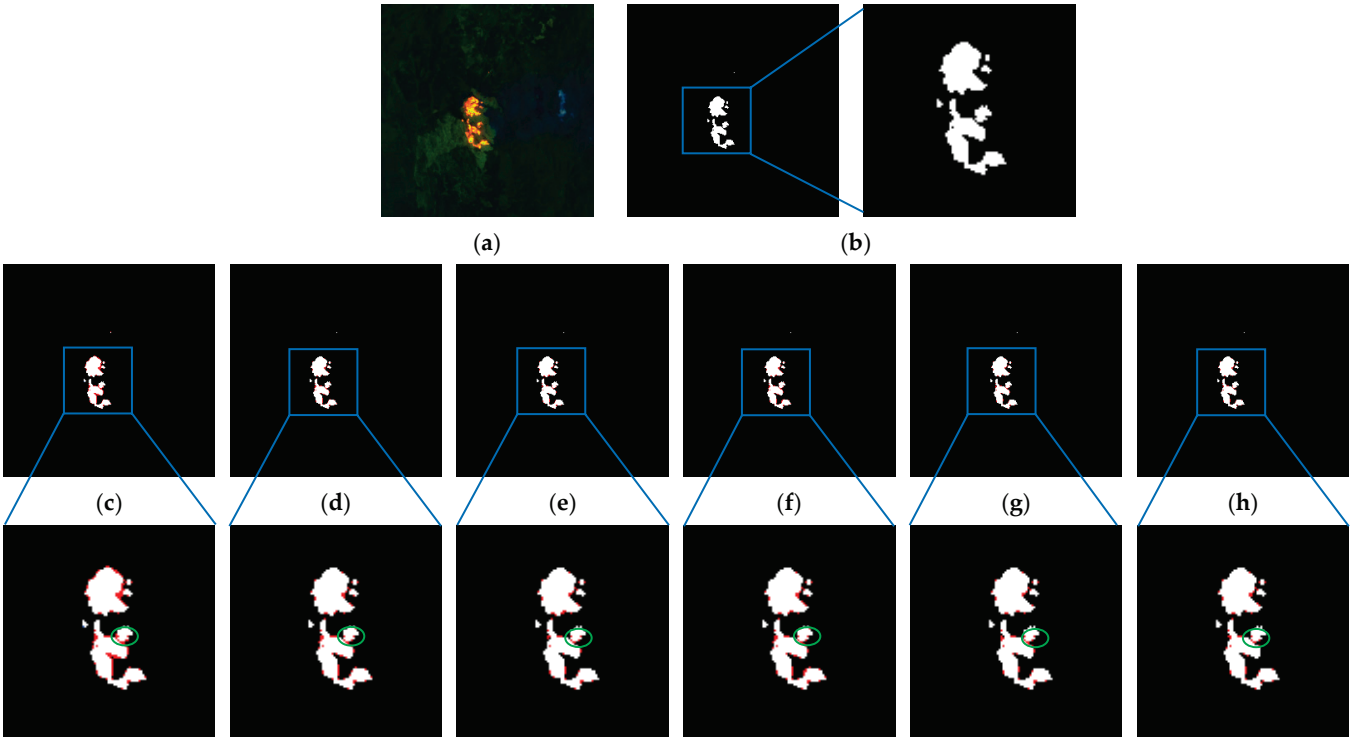


Figure 13. Segmentation results of image 1 by different CNN methods: (**a**) original image, (**b**) ground truth, (**c**) U-Net, (**d**) ResU-Net, (**e**) U-Net3+, (**f**) SwinU-Net, (**g**) WET-UNet, (**h**) FDE U-Net.

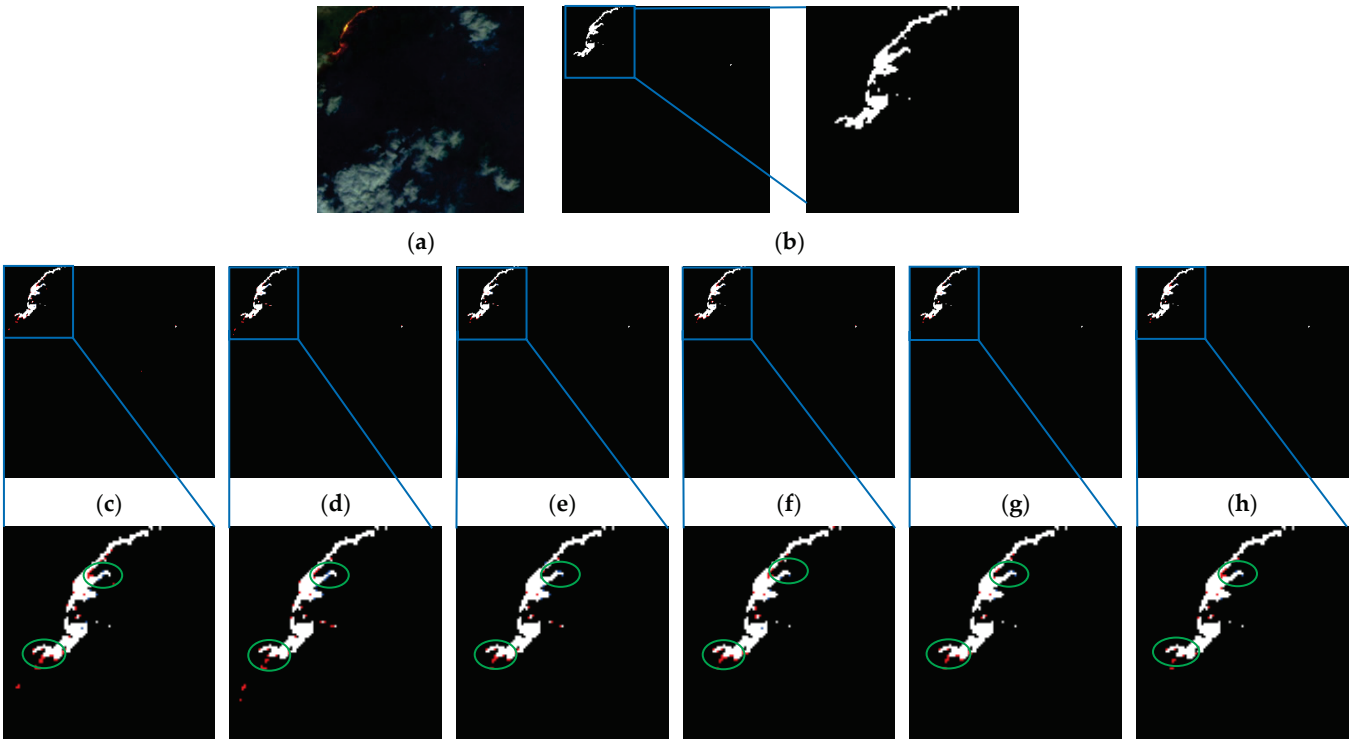


Figure 14. Segmentation results of image 2 by different CNN methods: (**a**) original image, (**b**) ground truth, (**c**) U-Net, (**d**) ResU-Net, (**e**) U-Net3+, (**f**) SwinU-Net, (**g**) WET-UNet, (**h**) FDE U-Net.

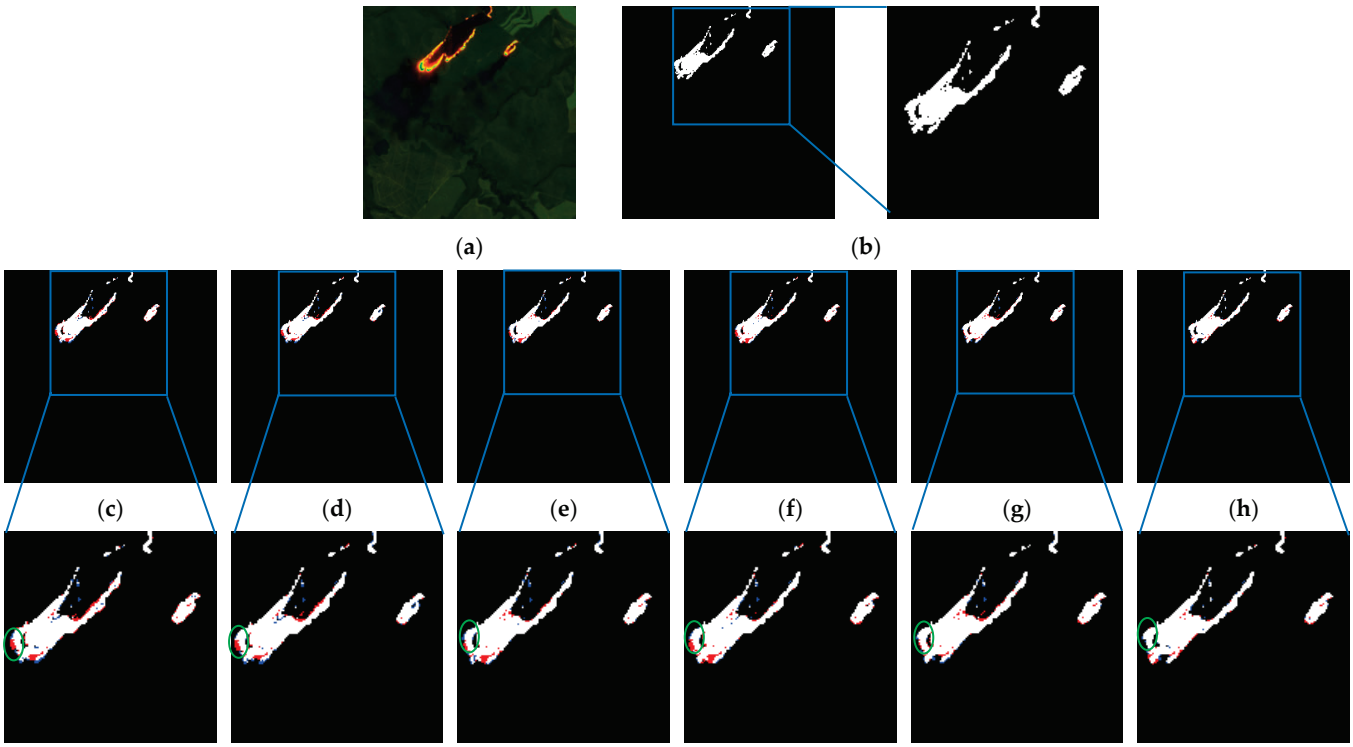


Figure 15. Segmentation results of image 3 by different CNN methods: (**a**) original image, (**b**) ground truth, (**c**) U-Net, (**d**) ResU-Net, (**e**) U-Net3+, (**f**) SwinU-Net, (**g**) WET-UNet, (**h**) FDE U-Net.

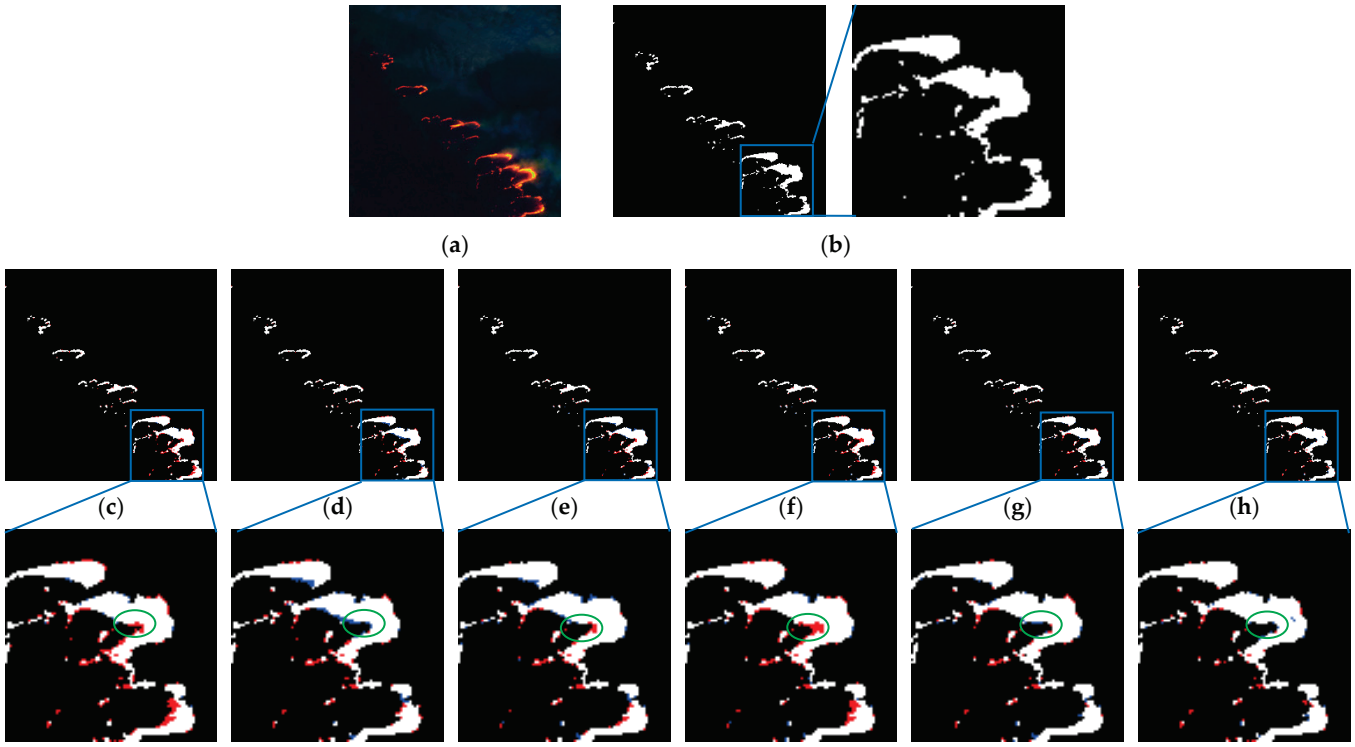


Figure 16. Segmentation results of image 4 by different CNN methods: (**a**) original image, **(b)** ground truth, (**c**) U-Net, (**d**) ResU-Net, (**e**) U-Net3+, (**f**) SwinU-Net, (**g**) WET-UNet, (**h**) FDE U-Net.

Table 3. Detection results of FDE U-Net and other CNN methods.

Method	Image 1				Image 2				Image 3				Image 4			
	GT	TP	FN	FP	GT	TP	FN	FP	GT	TP	FN	FP	GT	TP	FN	FP
U-Net	781	780	1	124	410	402	8	47	1347	1269	78	150	1711	**1693**	**18**	250
ResU-Net		781	0	74		401	9	36		1273	74	111		1632	79	146
U-Net3+		781	0	56		402	8	25		1281	66	95		1640	71	128
SwinU-Net		781	0	78		405	5	38		1300	47	121		1678	33	186
WET-UNet		781	0	53		404	6	25		1296	51	89		1664	47	106
FDE U-Net		**781**	**0**	**52**		**406**	**4**	**25**		**1306**	**41**	**68**		1683	28	**102**

As shown in Figures 13 and 14, the fire areas in the two scenarios are more concentrated, and most of the models are able to accurately find the exact location of the fire, the significant difference in the outputs is that the edges of the fire areas are not clearly segmented, such as in the elliptical box area in the figure, with more false positives as well as a small number of missed detections. The proposed method overcomes this problem by enhancing feature extraction, feature fusion, and detail processing. As shown in Figure 14c–g, this scene shows that there are many clouds in the image, and the fire area is relatively small. The high brightness of clouds easily confuses the fire area with the background area. The classic segmentation models U-Net, ResU-Net, U-Net3+, SwinU-Net, and WET-UNet generate a significant number of false positives, which are concentrated in the elliptical region in the lower-left corner of the fire area. As shown in Figure 14h, the numbers of false positives and false negatives are 25 and 4, respectively, which are the lowest among the compared methods. The area within the elliptical shows that the FDE U-Net performs the best boundary preservation for fire. Thus, the proposed method extracts rich, high-level semantic features and preserves the main information, which improves the network's detection ability in the case of a category imbalance. Irrelevant information and noise

are suppressed so that the model can effectively distinguish between fire boundaries and background pixels in cloudy scenes.

In Figures 15 and 16, the fire regions are irregular and scattered. In scene image 3, the segmentation results differ mainly around the edge of the fire region, within the elliptical box. This may be due to high-temperature fire pixels affecting the spectral reflectivity of nearby pixels through thermal radiation, causing misclassification. U-Net and ResU-Net have the highest false negatives, 78 and 74, respectively, with U-Net having more false positives. While U-Net3+ improves feature extraction through multi-scale fusion, reducing false positives to 95 and false negatives to 66, its contextual information fusion may be insufficient in complex or boundary-ambiguous scenarios. SwinU-Net combines the strengths of Swin Transformer to capture global information and long-range dependencies, making it particularly effective for detecting large fire regions. This leads to an improvement in recall and a reduction in false negatives to 47. However, it may still misidentify pixels in adjacent fire regions as part of the fire area, especially in edge regions where the boundaries are less clear, resulting in more false positives. WET-UNet takes advantage of wavelet transform in multi-scale analysis and combines it with the modeling capabilities of the transformer to achieve comprehensive extraction and refined processing of image features, optimizing the balance between precision and recall. The fire areas in the image 4 scene of Figure 16 are of different sizes and unevenly distributed, and there are separate fires detached from the main fire area that can be accurately detected by all models. The main difference lies in the densely distributed fire areas, such as the region within the elliptical box in Figure 16c–g, where U-Net, ResU-Net, U-Net3+, SwinU-Net, and WET-UNet all have varying negatives and false positives. In contrast, our method optimizes the capture and fusion of contextual information in the structural design and generates results closest to the ground truth. The results in Figures 15h and 16h show that both false positives and false negatives are minimal. It shows that FDE U-Net extracts and integrates information-rich high-level features from the input. By fusing different bands, the model's focus area is adjusted according to the response characteristics of different bands. The feature fusion and embedded attention mechanism enhance the model's ability to capture the fire area. Through wavelet transform, the low-frequency and high-frequency information extracted between multiple bands helps enhance the correlation between features in different bands, allowing the model to better utilize information from different bands for fire identification, pay more attention to the fire body and edge details, and, ultimately, achieve accurate segmentation of the fire area.

To further validate the generalization ability of our proposed model, we conduct additional experiments on the FLAME dataset with 2003 images, which is collected by drones during planned burns in the pine forests of Arizona. Figure 17 presents representative samples from the dataset, highlighting its key visual characteristics. The images in this dataset have higher spatial resolution and differ significantly from the Landsat-8 remote sensing images used in the main experiment in terms of data source and observation scale. The detection results on this dataset are shown in Table 4.

Table 4. Detection results on FLAME dataset.

Method	P (%)	R (%)	IoU (%)	F1 (%)
U-Net	89.03	84.94	76.89	86.94
ResU-Net	91.72	82.73	76.97	86.99
U-Net3+	90.07	87.10	79.46	88.56
SwinU-Net	92.61	88.84	83.04	90.72
WET-UNet	93.45	87.56	82.58	90.45
FDE U-Net	**93.82**	**89.73**	**84.69**	**91.71**



Figure 17. Examples from the FLAME dataset.

In the experiment on the FLAME dataset, the images are captured by close-range drones, which provide higher spatial resolution and more complex background texture information, such as ground vegetation, smoke occlusion, and buildings. These factors make it more difficult for the model to accurately extract pure fire area features, thus affecting detection accuracy. Table 4 shows the detection results on the FLAME dataset. As a result, the overall evaluation metrics are slightly lower than those on the Landsat-8 dataset. Despite this, the proposed FDE U-Net model still achieves the best detection performance on this dataset, demonstrating strong robustness and generalization ability. Specifically, the Precision, Recall, IoU, and F1 scores of FDE U-Net reach 93.82%, 89.73%, 84.69%, and 91.71%, respectively. Compared with other models, FDE U-Net shows the smallest performance drop. This observation indicates that the HWD, CBAM, and ACMix modules introduced into the model have good adaptability to changes in fire point distribution and target size.

4.3.2. Analysis of Small Targets and Cloud Cover Results

In order to further verify the role of the model in small target detection, we construct a dataset containing only small flame areas. We select 100 images with only small fires and apply horizontal and vertical flipping to generate a total of 300 small-fire images, several examples of which are shown in Figure 18. The experimental results show that the overall performance of the model decreases compared to the large-scale dataset. However, despite the overall more difficult task, FDE U-Net still performs better than other comparison models in this scenario, with an IoU of 87.89% and an F1 score of 93.49%, as shown in Table 5, maintaining its leading performance in small target recognition. This result indirectly proves that the edge detail preservation mechanism provided by the HWD module plays a key role in small-scale flame detection.

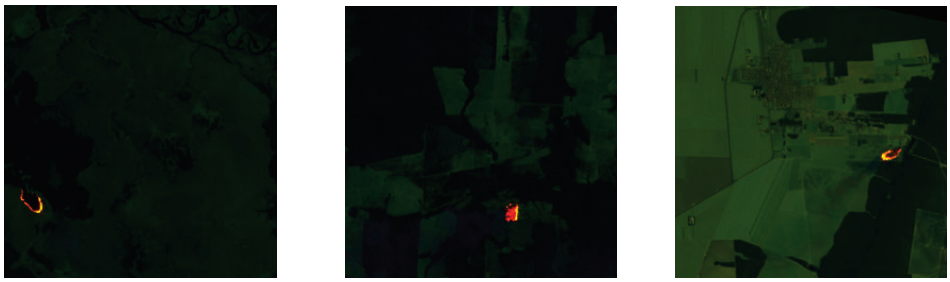


Figure 18. Small target scene examples.

Table 5. Detection results on small objects.

Method	P (%)	R (%)	IoU (%)	F1 (%)
U-Net	92.76	85.31	80.36	88.88
ResU-Net	93.18	84.07	79.55	88.39
U-Net3+	91.95	91.45	84.85	91.70
SwinU-Net	94.19	91.32	86.58	92.71
WET-UNet	94.35	90.66	86.36	92.62
FDE U-Net	**95.33**	**92.36**	**87.89**	**93.49**

Cloud cover also has an interfering effect on remote sensing fire detection tasks. Clouds block the spectral and thermal radiation characteristics of the fire source area, while introducing a large amount of invalid background information and destroying the edge clarity of the image, thus reducing the model's discrimination ability. To evaluate the robustness of each model under complex weather interference, we construct a dataset containing cloud-occluded scenes by increasing the proportion of cloud-containing images to 70%, several examples of which are shown in Figure 19. Table 6 shows the detection results on this dataset.

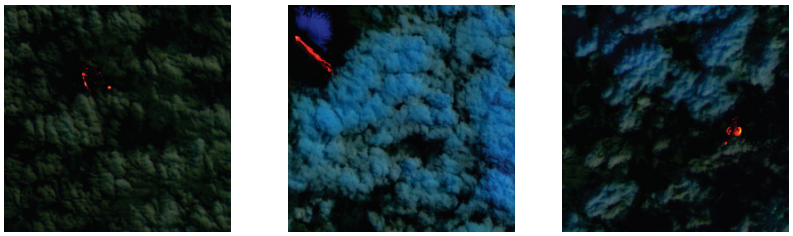


Figure 19. Cloud cover scene examples.

Table 6. Detection results with cloud cover.

Method	P (%)	R (%)	IoU (%)	F1 (%)
U-Net	93.52	84.21	79.89	88.64
ResU-Net	94.85	83.66	80.12	88.91
U-Net3+	95.28	86.17	82.87	90.50
SwinU-Net	96.21	91.31	88.00	93.59
WET-UNet	95.89	91.54	88.19	93.72
FDE U-Net	**96.41**	**92.06**	**88.98**	**94.18**

The experimental results show that all indicators of all models decrease to varying degrees under cloud interference. Despite this, FDE U-Net still achieves the best performance in precision, recall, IoU, and F1 score, demonstrating strong anti-interference capability. For example, the traditional U-Net model only reaches an intersection over union of 79.89, whereas FDE U-Net reaches 88.98, as shown in Table 6, representing an improvement of nearly nine percentage points. This indicates that its structure is more effective in maintaining edge information and suppressing background interference. This suggests that the model not only accurately detects fire regions but also effectively reduces the false positives and false negatives caused by clouds. These results show that FDE U-Net exhibits stronger robustness and generalization ability in complex environments such as cloud cover.

5. Discussion

In this study, we systematically evaluate the performance of the proposed FDE U-Net network for fire detection in remote sensing images. The experiments include both

quantitative and qualitative evaluation dimensions. In the quantitative evaluation, the performance of the model on fire detection is evaluated by a set of standardized metrics: Precision, Recall, IoU, and F1 scores. Through comparative analysis and ablation studies, we further confirm the effectiveness of the model, and compared with existing advanced methods, our proposed method performs better in all scenarios. To evaluate the efficiency of the proposed model, we compare the number of parameters, inference time, and FLOPS. It achieves faster inference time with a moderate number of parameters, which shows that the model achieves a good balance between performance and efficiency. In the qualitative evaluation, by showing the detection results of the model in different scenarios, it can be seen that the proposed model can accurately locate the fire area with minimal false positives and false negatives. In summary, FDE U-Net has excellent performance in fire detection tasks.

Although the overall performance of the model has been improved, there are still some limitations. Firstly, the innovation of FDE U-Net is mainly reflected in the combination and integration of existing modules. In the future, it is possible to consider building modules and feature extraction methods that can be used for fire detection based on the characteristics of remote sensing fire images. Secondly, relying on the traditional U-Net framework, it is still slightly insufficient in multi-scale modeling and cross-resolution feature fusion. Compared with emerging architectures based on transformer or multi-branch structures, the model may have information loss in processing complex fire forms. In the future, we will improve the innovation and practicality of the model from the following directions: firstly, it is necessary to explore new network backbone structures, such as transformer or Mamba, as well as dedicated extraction modules for fire features to enhance global perception capabilities; secondly, we can integrate multi-source remote sensing and ground data (such as MODIS, Sentinel, LiDAR, ground monitoring) to improve scene understanding capabilities; thirdly, we will develop lightweight models and semi-supervised learning methods to adapt to the needs of data scarcity and real-time detection.

6. Conclusions

We propose a novel FDE U-Net model for fire detection using Landsat-8 multispectral remote sensing data. The ACmix, CBAM, and HWD modules are integrated into each encoder layer in ResU-Net to extract the information-rich high-level features. ACmix combines convolution and self-attention mechanisms, where convolution effectively captures local features and self-attention dynamically captures global contextual information. By integrating the two, the model can handle local details while focusing on global features, enhancing its ability to understand complex images. CBAM uses channel and spatial attention mechanisms to focus on key areas in fire images by applying weights to features. This improves the model's ability to recognize essential regions and enhances the detection accuracy. The Haar wavelet downsampling technique can effectively capture detailed features, avoid the loss of important information during the downsampling process, and retain more detail and boundary information. The experimental results indicate that the proposed FDE U-Net performs exceptionally well in fire detection tasks, accurately extracting fire regions, even when the fire area is small and the background is complex. For the terrain interference, although the terrain will affect the distribution of background information of local images, such as vegetation occlusion and mountain shadows, the microscopic interference of the terrain will not significantly affect the model's recognition results of the fire area due to the spatial resolution of Landsat-8 of 30m. Therefore, in this study, terrain interference is not the main influencing factor, but we recognize that it may have an impact on the detection process, which can be further studied in the future.

Author Contributions: R.Z.: conceptualization, investigation, methodology, validation, visualization, formal analysis, writing (original draft). Z.X.: conceptualization, funding acquisition, investigation, resources, supervision, formal analysis, article revision. G.L. and P.H.: project administration, supervision, resources, formal analysis. R.W. and Y.Q.: conceptualization, methodology, validation, formal analysis. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Natural Science Foundation of China under Grant 62361058 and 62171272, Yunnan Expert Workstation under Grant 202305AF150012 (Corresponding author: Zhihui Xin).

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

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

References

1. Adams, M.A.; Shen, Z. Introduction to the characteristics, impacts and management of forest fire in China. *For. Ecol. Manag.* **2015**, *356*, 1. [CrossRef]
2. 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]
3. Global Wildfire Information System (2025)—With Minor Processing by Our World in Data. Available online: https://ourworldindata.org/grapher/annual-area-burnt-by-wildfires?tab=maps (accessed on 22 June 2025).
4. Abdusalomov, A.; Umirzakova, S.; Bakhtiyor Shukhratovich, M.; Mukhiddinov, M.; Kakhorov, A.; Buriboev, A.; Jeon, H.S. Drone-Based Wildfire Detection with Multi-Sensor Integration. *Remote Sens.* **2024**, *16*, 4651. [CrossRef]
5. Fraser, R.H.; Van der Sluijs, J.; Hall, R.J. Calibrating Satellite-Based Indices of Burn Severity from UAV-Derived Metrics of a Burned Boreal Forest in NWT, Canada. *Remote Sens.* **2017**, *9*, 279. [CrossRef]
6. Tong, H.; Yuan, J.; Zhang, J.; Wang, H.; Li, T. Real-Time Wildfire Monitoring Using Low-Altitude Remote Sensing Imagery. *Remote Sens.* **2024**, *16*, 2827. [CrossRef]
7. Thangavel, K.; Spiller, D.; Sabatini, R.; Amici, S.; Sasidharan, S.T.; Fayek, H.; Marzocca, P. Autonomous satellite wildfire detection using hyperspectral imagery and neural networks: A case study on Australian wildfire. *Remote Sens.* **2023**, *15*, 720. [CrossRef]
8. Sun, Y.; Jiang, L.; Pan, J.; Sheng, S.; Hao, L. A satellite imagery smoke detection framework based on the Mahalanobis distance for early fire identification and positioning. *Int. J. Appl. Earth Obs. Geoinf.* **2023**, *118*, 103257. [CrossRef]
9. Barmpoutis, P.; Papaioannou, P.; Dimitropoulos, K.; Grammalidis, N. A Review on Early Forest Fire Detection Systems Using Optical Remote Sensing. *Sensors* **2020**, *20*, 6442. [CrossRef]
10. Pourshakouri, F.; Darvishsefat, A.A.; Samadzadegan, F.; Attarod, P.; Amini, S. An improved algorithm for small and low-intensity fire detection in the temperate deciduous forests using MODIS data: A preliminary study in the Caspian Forests of Northern Iran. *Nat. Hazards* **2023**, *116*, 2529–2547. [CrossRef]
11. Cardíl, A.; Tapia, V.M.; Monedero, S.; Quiñones, T.; Little, K.; Stoof, C.R.; Ramirez, J.; de Miguel, S. Characterizing the rate of spread of large wildfires in emerging fire environments of northwestern Europe using Visible Infrared Imaging Radiometer Suite active fire data. *Nat. Hazards Earth Syst. Sci.* **2023**, *23*, 361–373. [CrossRef]
12. Zhang, T.; Wooster, M.J.; Xu, W. Approaches for synergistically exploiting VIIRS I- and M-Band data in regional active fire detection and FRP assessment: A demonstration with respect to agricultural residue burning in Eastern China. *Remote Sens. Environ.* **2017**, *198*, 407–424. [CrossRef]
13. Hong, Z.; Tang, Z.; Pan, H.; Zhang, Y.; Zheng, Z.; Zhou, R.; Ma, Z.; Zhang, Y.; Han, Y.; Wang, J.; et al. Active fire detection using a novel convolutional neural network based on Himawari-8 satellite images. *Front. Environ. Sci.* **2022**, *10*, 794028. [CrossRef]
14. Thompson, D.K.; Morrison, K. A classification scheme to determine wildfires from the satellite record in the cool grasslands of southern Canada: Considerations for fire occurrence modelling and warning criteria. *Nat. Hazards Earth Syst. Sci.* **2020**, *20*, 3439–3454. [CrossRef]
15. Ghali, R.; Jmal, M.; Mseddi, W.S.; Attia, R. Recent advances in fire detection and monitoring systems: A review. In Proceedings of the 10th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications, Hammamet, Tunisia, 20–22 December 2020. [CrossRef]
16. Zhang, Z.; Shen, T.; Zou, J. An improved probabilistic approach for fire detection in videos. *Fire Technol.* **2014**, *50*, 745–752. [CrossRef]
17. Celik, T.; Demirel, H. Fire detection in video sequences using a generic color model. *Fire Saf. J.* **2009**, *44*, 147–158. [CrossRef]
18. Chen, X.; An, Q.; Yu, K.; Ban, Y. A novel fire identification algorithm based on improved color segmentation and enhanced feature data. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 3075380. [CrossRef]

19. Jang, E.; Kang, Y.; Im, J.; Lee, D.-W.; Yoon, J.; Kim, S.-K. Detection and monitoring of forest fires using Himawari-8 geostationary satellite data in South Korea. *Remote Sens.* **2019**, *11*, 271. [CrossRef]
20. Tennant, E.; Jenkins, S.F.; Miller, V.; Robertson, R.; Wen, B.; Yun, S.-H.; Taisne, B. Automating tephra fall building damage assessment using deep learning. *Nat. Hazards Earth Syst. Sci.* **2024**, *24*, 4585–4608. [CrossRef]
21. Zhu, J.; Zhang, H.; Li, S.; Wang, S.; Ma, H. Cross teaching-enhanced multispectral remote sensing object detection with transformer. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 2401–2413. [CrossRef]
22. Zhang, S.; Zhang, S.; Guo, T.; Xu, R.; Liu, Z.; Du, Q. Clutter Modeling and Characteristics Analysis for GEO Spaceborne-Airborne Bistatic Radar. *Remote Sens.* **2025**, *17*, 1222. [CrossRef]
23. Yu, L.; Qin, H.; Huang, S.; Wei, W.; Jiang, H.; Mu, L. Quantitative study of storm surge risk assessment in an undeveloped coastal area of China based on deep learning and geographic information system techniques: A case study of Double Moon Bay. *Nat. Hazards Earth Syst. Sci.* **2024**, *24*, 2003–2024. [CrossRef]
24. Hwang, G.; Jeong, J.; Lee, S.J. SFA-Net: Semantic Feature Adjustment Network for Remote Sensing Image Segmentation. *Remote Sens.* **2024**, *16*, 3278. [CrossRef]
25. Kukuk, S.B.; Kilimci, Z.H. Comprehensive analysis of forest fire detection using deep learning models and conventional machine learning algorithms. *Int. J. Comput. Exp. Sci. Eng.* **2021**, *7*, 84–94. [CrossRef]
26. Jin, C.; Wang, T.; Alhusaini, N.; Zhao, S.; Liu, H.; Xu, K.; Zhang, J. Video fire detection methods based on deep learning: Datasets, methods, and future directions. *Fire* **2023**, *6*, 315. [CrossRef]
27. Sun, H.R.; Shi, B.J.; Zhou, Y.T.; Chen, J.H.; Hu, Y.L. A smoke detection algorithm based on improved YOLO v7 lightweight model for UAV optical sensors. *IEEE Sens. J.* **2024**, *24*, 26136–26147. [CrossRef]
28. Hu, X.; Liu, W.; Wen, H.; Yuen, K.-V.; Jin, T.; Junior, A.C.N.; Zhong, P. AF-Net: An Active Fire Detection Model Using Improved Object-Contextual Representations on Unbalanced UAV Datasets. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 13558–13569. [CrossRef]
29. 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]
30. Han, Y.; Zheng, C.; Liu, X.; Tian, Y.; Dong, Z. Burned Area and Burn Severity Mapping with a Transformer-Based Change Detection Model. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 13866–13880. [CrossRef]
31. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015, Munich, Germany, 5–9 October 2015. [CrossRef]
32. Siddique, N.; Paheding, S.; Elkin, C.P.; Devabhaktuni, V. U-Net and its variants for medical image segmentation: A review of theory and applications. *IEEE Access* **2021**, *9*, 82031–82057. [CrossRef]
33. Zhang, P.; Ban, Y.; Nascetti, A. Learning U-Net without forgetting for near real-time wildfire monitoring by the fusion of SAR and optical time series. *Remote Sens. Environ.* **2021**, *261*, 112467. [CrossRef]
34. Xu, Z.; Li, H.; Long, B. MSU-Net: Multi-scale self-attention semantic segmentation method for oil-tea camellia planting area extraction in hilly areas of southern China. *Expert Syst. Appl.* **2025**, *263*, 125779. [CrossRef]
35. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015. [CrossRef]
36. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017. [CrossRef]
37. Wang, Z.; Yang, P.; Liang, H.; Zheng, C.; Yin, J.; Tian, Y.; Cui, W. Semantic segmentation and analysis on sensitive parameters of forest fire smoke using smoke-unet and landsat-8 imagery. *Remote Sens.* **2021**, *14*, 45. [CrossRef]
38. Afsar, R.; Sultana, A.; Abouzahra, S.N.; Aspiras, T.; Asari, V.K. Using ResWnet for semantic segmentation of active wildfires from Landsat-8 imagery. In Proceedings of the SPIE—The International Society for Optical Engineering, National Harbor, MD, USA, 19–24 August 2024. [CrossRef]
39. Schroeder, W.; Oliva, P.; Giglio, L.; Quayle, B.; Lorenz, E.; Morelli, F. Active fire detection using Landsat-8/OLI data. *Remote Sens. Environ.* **2016**, *185*, 210–220. [CrossRef]
40. Kumar, S.S.; Roy, D.P. Global operational land imager Landsat-8 reflectance-based active fire detection algorithm. *Int. J. Digit. Earth* **2018**, *11*, 154–178. [CrossRef]
41. Murphy, S.W.; de Souza Filho, C.R.; Wright, R.; Sabatino, G.; Pabon, R.C. HOTMAP: Global hot target detection at moderate spatial resolution. *Remote Sens. Environ.* **2016**, *177*, 78–88. [CrossRef]
42. de Almeida Pereira, G.H.; Fusioka, A.M.; Nassu, B.T. Active fire detection in Landsat-8 imagery: A large-scale dataset and a deep-learning study. *ISPRS J. Photogramm. Remote Sens.* **2021**, *178*, 171–186. [CrossRef]
43. Pan, X.; Ge, C.; Lu, R.; Song, S.; Chen, G.; Huang, Z.; Huang, G. On the Integration of Self-Attention and Convolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 805–815. [CrossRef]

44. 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. [CrossRef]
45. Xu, G.; Liao, W.; Zhang, X.; Li, C.; He, X.; Wu, X. Haar wavelet downsampling: A simple but effective downsampling module for semantic segmentation. *Pattern Recognit.* **2023**, *143*, 109819. [CrossRef]
46. Zhang, Z.; Liu, Q.; Wang, Y. Road extraction by deep residual u-net. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 749–753. [CrossRef]
47. Huang, H.; Lin, L.; Tong, R.; Hu, H.; Zhang, Q.; Iwamoto, Y. Unet 3+: A full-scale connected unet for medical image segmentation. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Barcelona, Spain, 4–8 May 2020. [CrossRef]
48. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Swin-Unet: Unet-Like Pure Transformer for Medical Image Segmentation. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022. [CrossRef]
49. Zeng, Y.; Li, J.; Zhao, Z.; Liang, W.; Zeng, P.; Shen, S.; Zhang, K.; Shen, C. WET-UNet: Wavelet integrated efficient transformer networks for nasopharyngeal carcinoma tumor segmentation. *Sci. Prog.* **2024**, *107*, 00368504241232537. [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.

fire

MDPI

Article

Multi-Scale Construction Site Fire Detection Algorithm with Integrated Attention Mechanism

Haipeng Sun [1] and Tao Yao [1,2,*]

[1] College of Information Science and Technology, Southwest Jiaotong University, Chengdu 610097, China; heiven@my.swjtu.edu.cn
[2] Yantai New Generation Information Technology Research Institute, Southwest Jiaotong University, Yantai 264000, China
* Correspondence: yaotao@ldu.edu.cn

Abstract: The occurrence of construction site fires is consistently accompanied by casualties and property damage. To address the issues of large target-scale variations and frequent false detections in construction site fire monitoring, we propose a fire detection algorithm based on an improved YOLOv8 model, achieving real-time and efficient detection of fires on construction sites. First, considering the wide range of scale variations in detected objects, an additional detection layer with a 64-times down-sampling rate is introduced to enhance the algorithm's detection capability for multi-scale targets. Then, the MBConv module and the ESE attention block are integrated into the C2f structure to enhance feature extraction capabilities while reducing computational complexity. An iCBAM attention module is designed to suppress background noise interference and enhance the representation capability of the network. Finally, the WIoUv3 metric is adopted in the loss function for bounding box regression to mitigate harmful gradient issues. Comparative experiments demonstrate that, on a self-constructed construction site fire dataset, the improved algorithm achieves an accuracy and recall increase of 4.6% and 3.0%, respectively, compared to the original YOLOv8 model. Additionally, mAP50 and mAP50-95 are improved by 1.6% and 1.5%, respectively. This algorithm provides a more effective solution for fire monitoring in construction environments.

Keywords: YOLOv8; multi-scale detection; fire detection; attention mechanism; Wise-IoU

1. Introduction

With the rapid advancement of global modernization and infrastructure development, the scale of engineering projects continues to expand, significantly increasing the complexity of construction environments and posing severe challenges to construction site safety management. According to data from China's National Fire and Rescue Administration, in the first half of 2023, a total of 550,000 fires were reported nationwide, resulting in 959 fatalities and direct property losses of 3.94 billion yuan, demonstrating a clear upward trend compared to the same period last year. As losses caused by fires exhibit an exponential positive correlation with response time, timely detection and extinguishment of fire sources at their initial stage can greatly reduce damage [1]. Therefore, there is an urgent need for a real-time and accurate fire detection solution.

Traditional fire monitoring methods mainly rely on manual inspections and sensor networks [2]. Manual inspection suffers from poor real-time performance and high missed detection rates, whereas sensor networks are susceptible to environmental interference (such as dust and illumination variations) and entail high deployment costs. To overcome

the limitations of traditional monitoring methods and enhance supervisory efficiency, object detection algorithms based on deep learning have provided new approaches for fire detection in recent years [3].

Two-stage object detection algorithms have attracted considerable attention in the field of fire detection due to their superior detection accuracy and robust feature representation capabilities. Choi [4] et al. integrated the Swin Transformer backbone into the Faster R-CNN framework, achieving excellent performance in multi-category wildfire detection tasks. This method notably enhanced the identification capability for small-scale flame and smoke targets and effectively reduced the false detection rate between smoke and non-fire smoke, attaining an mAP of 0.841 and demonstrating strong robustness. Cheknane [5] et al. proposed an improved Faster R-CNN model incorporating a hybrid feature extraction mechanism to address the challenges of flame and smoke detection in complex indoor and outdoor scenarios. Their method achieved outstanding performance, with an mAP of 90.1% and an accuracy of 96.5% on a private dataset. Through multi-level feature fusion, the model significantly improved its perception of diverse fire scenarios. However, two-stage object detection algorithms typically suffer from substantial model parameter sizes and relatively slow inference speeds, making it particularly challenging to meet the real-time requirements of fire warning systems. Consequently, these methods have not yet been widely implemented in practical engineering applications.

Single-stage algorithms have achieved numerous breakthroughs in the field of fire detection. Gragnaniello [6] et al. proposed a detection framework named "FLAME," whose innovation lies in exploiting flame motion characteristics to filter out background interference and enhance detection accuracy. Experimental results indicated that, compared with the original YOLO model, incorporating the motion analysis module significantly reduced the false alarm rate, increasing the accuracy and F-score by 18% and 10%, respectively, with an average alarm delay of 9.17 s, thus satisfying real-time alarm requirements. Mukhriddinov [7] et al. proposed a lightweight, improved YOLOv5 algorithm aimed at early smoke detection for outdoor wildfire scenarios. By re-optimizing anchor box dimensions and introducing a fast spatial pyramid pooling module (SPP-Fast) to strengthen small-scale feature extraction, alongside employing a bidirectional feature pyramid network BiFPN to enhance multi-scale feature fusion efficiency, combined with model pruning and transfer learning, the adaptability of the model in resource-constrained environments was significantly improved. Experimental results on an unmanned aerial vehicle (UAV)-captured wildfire smoke dataset demonstrated that the proposed method achieved an average precision of 73.6%, exhibiting promising real-time detection performance. Liu [8] et al. proposed a detection network termed TFNet, integrating convolutional neural networks and Transformer architectures, to address the challenging problem of identifying small-scale fire spots and sparse smoke in the early detection of forest fires. The proposed method introduces a multi-branch scale-aware representation S-R module, effectively enhancing multi-scale feature representation capabilities. Additionally, a Context-Guided Multi-Scale Feature Fusion CG-MSFF encoder is developed to efficiently integrate local and global contextual information. Meanwhile, a Transformer-based decoding detection head combined with a Weighted IoU loss function is employed, significantly improving the model's ability to identify and localize challenging samples. TFNet achieved superior detection performance compared to baseline models on two publicly available fire detection datasets. Li [9] et al. proposed an improved algorithm, YOLO11s-MSCA, by integrating a Multi-Scale Convolutional Attention MSCA module with the YOLO11 framework to address challenges associated with variations in fire target scales and complex environmental background interference. Experimental results demonstrated that this method achieved superior detection accuracy and generalization performance across multiple publicly available datasets, significantly

improving the detection precision for small-scale fire and smoke, while maintaining the high efficiency and low complexity required for real-time detection.

Although deep learning algorithms have made notable breakthroughs in the field of fire detection, existing methods still exhibit deficiencies in the collaborative optimization of multi-scale detection efficiency and robustness against complex backgrounds. To address these challenges, this paper proposes an improved YOLOv8-based algorithm named YOLO-Fire, with key innovations summarized as follows:

(1) Self-Constructed Construction Site Fire Dataset:

Current mainstream object detection datasets lack annotation data specifically targeted at smoke and fire in fire scenarios, and there is currently no authoritative and publicly available specialized dataset for construction site fire detection. To bridge this gap, this paper constructs a self-developed construction site fire detection dataset (Fire-5000) through multi-channel collection, precise annotation, and data augmentation.

(2) Multi-scale Lightweight Collaborative Architecture:

An additional detection layer with a 64-times down-sampling rate is introduced to broaden the detection scale coverage. Meanwhile, a lightweight convolutional module, termed Efficient Mobile Bottleneck (EMB), is designed and integrated into the C2f unit, mitigating the computational overhead caused by increased network depth and achieving a balanced trade-off between detection efficiency and accuracy.

(3) Inverted Convolutional Block Attention Module (iCBAM):

By embedding the Convolutional Block Attention Module (CBAM) within the Inverted Residual Block (IRB), we guide the network to selectively focus on informative features across both channel and spatial dimensions, which leads to a more powerful feature representation.

(4) Dynamic Gradient Optimization Strategy:

The WIoU v3 loss function is introduced, which dynamically adjusts gradient weights through adaptive assessment of sample outliers. This effectively mitigates interference caused by extreme-quality samples during model training, thereby enhancing the stability and convergence efficiency of the training process.

2. Materials and Methods

2.1. YOLOv8 Algorithm

The YOLO series of algorithms was first introduced by Joseph Redmon [10] et al. in their 2015 paper, innovatively formulating the object detection task as a regression problem within a single convolutional neural network. This approach effectively achieves a remarkable balance between real-time performance and detection accuracy. In subsequent developments, the YOLO algorithm has undergone several significant iterations and optimizations, each yielding substantial improvements in detection accuracy, inference speed, and model lightweighting. Among these versions, the YOLOv5 [11] model developed by Ultralytics has gained widespread recognition and adoption in industry due to its efficiency, stability, and ease of deployment. YOLOv8 [12], inheriting the efficient architecture of YOLOv5, achieves further architectural majorization and technological innovations. Its network structure primarily consists of the following core components:

(1) Backbone: The backbone network is built upon an improved CSPDarkNet architecture, employing stacked convolutional layers and bottleneck structures to facilitate efficient feature extraction. Compared to YOLOv5, the key improvement in YOLOv8 is the replacement of the fundamental C3 module with the more efficient C2f module. Inspired by the ELAN [13] design, the new module introduces additional parallel gradient-flow branches, effectively reducing computational overhead while capturing richer gradient-flow information. Moreover, its basic convolutional unit adopts the CBS structure, con-

sisting of a two-dimensional convolutional layer, batch normalization, and the SiLU [14] activation function.

(2) Neck: The neck network adopts the FPN [15] (Feature Pyramid Network) combined with PAN [16] (Path Aggregation Network) structures, enabling efficient multi-scale feature fusion through a bidirectional feature pyramid approach.

(3) Detection Head: The detection layers perform detection tasks across multiple feature maps at different scales. The detection head employs a decoupled head design, separating classification and regression tasks to prevent interference between them, thereby enhancing the accuracy of object recognition.

2.2. Fire-5000 Dataset

Currently, in the field of construction site fire detection, there is no authoritative and publicly available dataset. Mainstream fire datasets typically include various fire scenarios without specific distinctions for work site images, lacking scenes that encompass distinctive corresponding characteristics. To address practical engineering needs, this paper constructs a custom construction site fire CV dataset. The dataset constructed in this study strictly follows a four-stage construction paradigm, namely: multi-source heterogeneous data acquisition and scene adaptation, content-consistency-based data cleaning, refined target labeling, and physically constrained data augmentation. The detailed implementation process is as follows:

(1) Data Collection and Screening: Fire-related datasets were collected from multiple publicly available dataset platforms, from which images specifically limited to construction site fire scenarios were selected. Relevant video materials were downloaded, and key frames were extracted using frame sampling techniques (one frame per second) to retain temporal feature images covering the entire developmental cycle of fires. The image collection contains construction site fire images from various fire development stages, lighting conditions, and resolutions, closely aligning with real application scenarios to enhance the specificity of feature extraction by the algorithm.

(2) Data Cleaning: During the integration of fire images, it was observed that a significant number of duplicate images related to construction site scenarios existed across different datasets. Using these repetitive samples during model training could cause severe problems, such as model overfitting and evaluation bias. To prevent the aforementioned issues, a two-stage data cleaning method was adopted in this study. Firstly, a Python script was developed to sequentially read all images from the directory, generating hash values for each image file using the MD5 algorithm. Identical hash values indicate duplicate image content and batch deleting of identical images. After the initial cleaning stage, images with identical visual content but differing in file format or resolution remained within the directory. Since such images produce largely similar features during feature extraction, they should also be considered duplicates and removed accordingly. Subsequently, a secondary cleaning process was conducted using the Structural Similarity (SSIM) algorithm based on the OpenCV library. The SSIM [17] index measures image similarity across three dimensions: contrast, luminance, and structure. In this study, a threshold $\theta \geq 0.6$ was set as the criterion for determining visual duplicates and then deleting similar files.

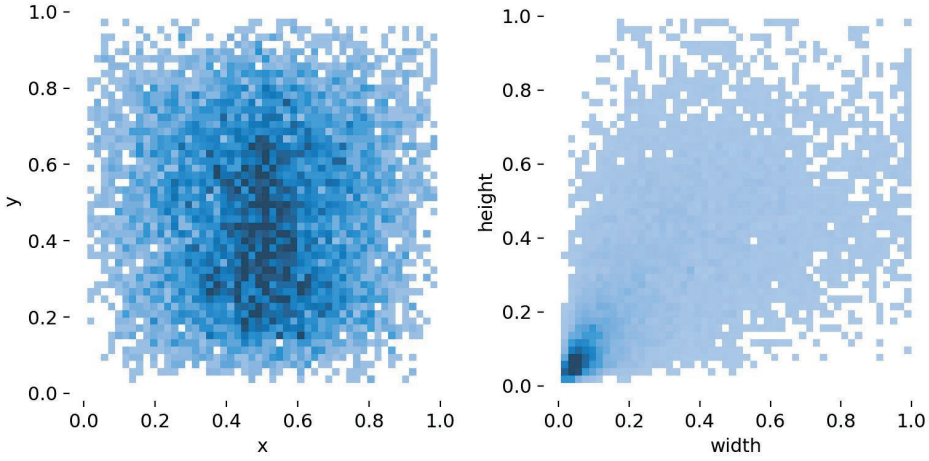
(3) Data labeling: Images of fire and smoke within the dataset were labeled using the LabelMe software. As Figure 1, the Minimum Bounding Rectangle strategy was employed, where fire was enclosed using bounding boxes, tightly fitting the full fire contours. Each bounding box was labeled with the target category “Fire”, and smoke was annotated similarly. After annotation completion, labels were batch exported and converted into the TXT format compatible with the YOLO model.



Figure 1. Annotating fire images using LabelMe software.

(4) Data Augmentation: To increase the number of samples, 500 images were selected from the annotated dataset and randomly processed using data augmentation techniques. The augmented images were then integrated into the dataset.

According to statistical analysis, the self-constructed dataset comprises a total of 5004 fire images, with approximately 10,500 annotated instances. Among these, there are approximately 6800 fire instances and 3700 smoke instances. The number of fire instances is nearly twice that of smoke instances, exhibiting a clear class imbalance. In real fire scenarios, fire tend to spread to surrounding combustible objects, forming multiple fire targets. In contrast, smoke generally has lower density, rises due to heat, and often aggregates into a single smoke region. Thus, the observed phenomenon of abundant fire targets and fewer smoke targets in the actual image collection aligns with authentic fire conditions, validating the reliability of the self-constructed dataset. As shown in Figure 2, color intensity is used to indicate the quantitative distribution of targets that meet the specified criteria. In the dataset, the distribution of target center points is relatively dispersed, with higher density in the central region, indicating a clear center-oriented composition tendency during data collection. Meanwhile, peripheral regions also exhibit a relatively uniform distribution, consistent with the realistic scenarios of fire occurrences. Furthermore, the scale distribution of targets exhibits clear multi-scale characteristics. Given the explicitly demonstrated multi-scale characteristics of targets within this dataset, it becomes particularly essential to conduct research focused on multi-scale optimization of object detection algorithms.



(**a**) Center Point Distribution (**b**) Width-Height Scale Distribution

Figure 2. Scale distribution of dataset targets: In (**a**), the *X* and *Y* axes represent the relative horizontal and vertical positions; In (**b**), the *X* and *Y* axes represent the relative width and height.

3. Algorithm Optimization

To address the problems of large-scale variations in target sizes and susceptible to environmental interference causing false detections in construction site fire scenarios, this paper optimizes the algorithm from three aspects: multi-scale detection [18], attention mechanisms, and the loss function, aiming to improve the algorithm's detection performance specifically in construction site contexts.

3.1. Multi-Scale Lightweight Collaborative Architecture

3.1.1. Multi-Scale Network Structure Optimization

The primary method for evaluating the performance of object detection algorithms is to compare their detection capabilities on comprehensive benchmark datasets, such as MS COCO. Such datasets typically encompass target objects of moderate size and exhibit relatively balanced scale distributions. Therefore, mainstream multi-scale detection algorithms, including the YOLO series, typically adopt a compromise strategy regarding structural parameters such as network depth, down-sampling rate, and feature fusion scales, aiming to achieve a balance in detection performance across targets of various scales. This strategy performs well in conventional object detection tasks; however, it exhibits notable limitations in specific application scenarios (such as the prevalence of large-scale targets in construction environments or small-object detection in remote sensing), making it challenging for the algorithm to effectively identify targets beyond the normal scale range [19]. As illustrated in Figure 3, the original YOLOv8 model exhibits missed detections for large-scale smoke targets and insufficient confidence (only 0.32) for small-scale fire targets during predictions on construction site fire images.

The official pretrained weights for the YOLOv8 algorithm were trained and optimized using the MS COCO [20] dataset, which covers 80 common object categories. Among these, approximately 72% of target instances have sizes concentrated between 32 $\times$ 32 and 256 $\times$ 256 pixels. Therefore, the original YOLOv8 network architecture incorporates only three scale levels with down-sampling rates of 32$\times$, 16$\times$, and 8$\times$, corresponding to theoretical receptive fields of 32 $\times$ 32, 16 $\times$ 16, and 8 $\times$ 8 pixels, respectively. While this structural design demonstrates good detection performance for medium-scale targets, it exhibits clear limitations when applied to special scenarios such as construction site fire incidents. In this paper, the network architecture is redesigned to enhance multi-scale detection performance.



(**a**) Missed detections of large-scale smoke regions. (**b**) The confidence score for small-scale fire is only 0.32.

Figure 3. Multi-scale performance of the YOLOv8 model in construction site fire scenarios.

According to the principle of down-sampling, after down-sampling operations, the resolution of feature maps gradually decreases, resulting in the loss of detailed geometric information within the images. However, this simultaneously expands the receptive field of model, enabling the network to capture richer semantic information. Therefore, increasing the down-sampling rate of the model can yield feature maps with larger receptive fields, thereby enhancing the model's capability to detect large-scale targets. To address the insufficient semantic information expressed by the original YOLOv8 architecture at its deepest output feature map with a 32-fold down-sampling rate, this study introduces an additional cascaded down-sampling pathway at the end of the backbone network. This added pathway consists of convolutional layers and the C2f module, further down-sampling to a 64× scale and generating a 10 × 10 feature map. Considering the real-time requirements of construction site fire detection tasks, the network width was adjusted after introducing the new down-sampling layer. Specifically, the number of channels in the original 32× down-sample feature map was reduced from 1024 to 768, while the newly introduced 64× down-sample feature map was set to 1024 channels, thus balancing detection accuracy and computational efficiency of the network, as presented in Figure 4.

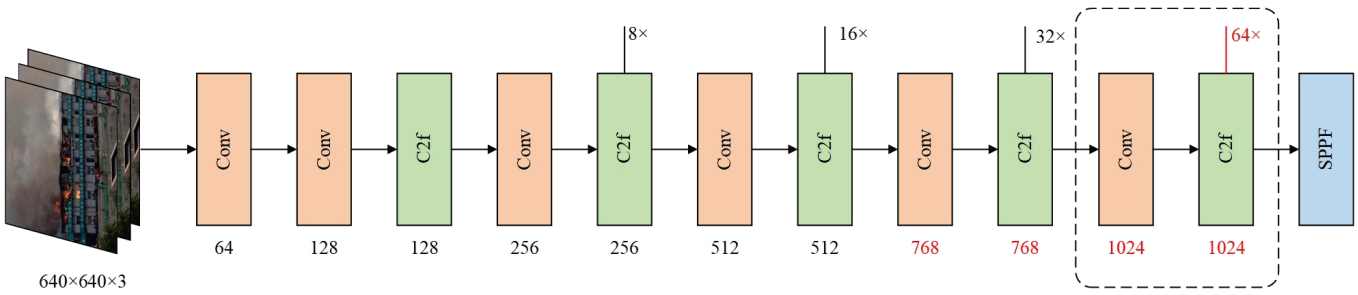


Figure 4. Expansion of Deep Feature Extraction Layers.

Meanwhile, to effectively utilize the newly introduced 64× down-sampling layer's feature information, this paper further designs an additional set of feature fusion pathways based on the original feature pyramid structure proposed by Ultralytics. This includes additional up-sampling and down-sampling operations, along with the corresponding construction of a detection head adapted specifically to this scale. The specific improvements are highlighted in green in Figure 5.

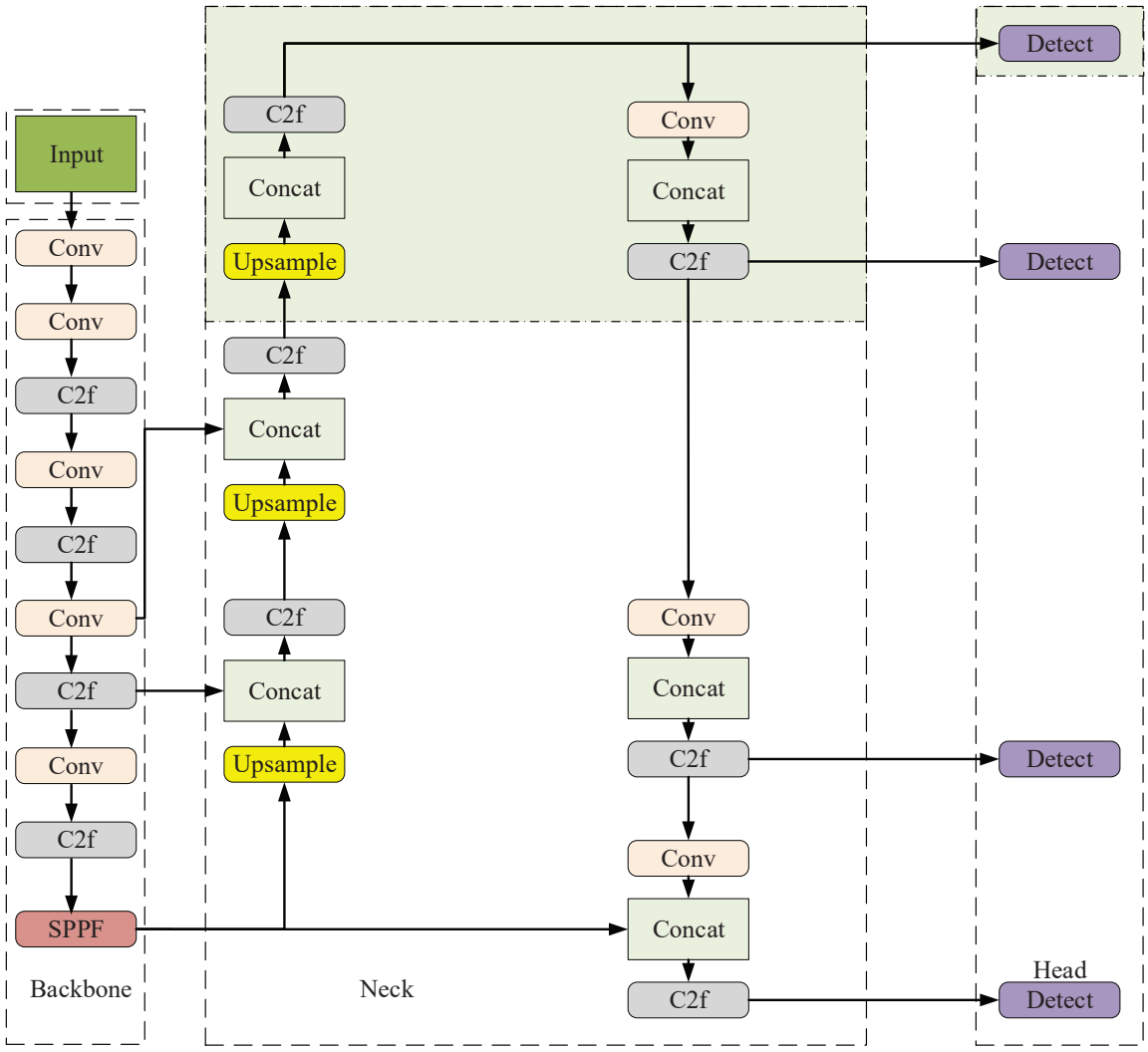


Figure 5. Network structure after multi-scale optimization.

3.1.2. Lightweight Convolution-Based Reconstruction of the C2f Module

The newly introduced 64× down-sampling detection layer described in the previous section significantly enhances the detection performance for large-scale targets. However, this addition increases the depth of the neural network, substantially elevating the computational complexity of the model, thus posing challenges to real-time construction site fire detection [21]. To address this issue, a lightweight collaborative convolutional block suitable for multi-scale detection is designed to reconstruct the C2f module, thereby balancing the efficiency and accuracy of the model.

Google's research team proposed the EfficientNet [22] architecture, a lightweight convolutional neural network that achieves higher detection accuracy with lower computational costs and fewer model parameters by hierarchically stacking MBConv [23] modules. The module's design is primarily inspired by depthwise separable convolutions and bottleneck structures, replacing conventional convolutions with a combination of pointwise convolutions and depthwise convolutions for feature extraction, significantly reducing computational complexity and the number of parameters.

For a convolution operation, suppose the input feature map has dimensions $W_i \times H_i$ with C_i channels, the convolution kernel size is $K_w \times K_h$, also with C_i channels, and the output feature map has dimensions $W_o \times H_o \times C_o$. In standard convolution, each kernel performs element-wise multiplication with every channel of the input and sums the

results, producing a scalar value at each spatial location. The computational complexity is represented by Equation (1):

$$\text{FLOPs}_{\text{Conv}} = W_o \times H_o \times C_o \times C_i \times K_w \times K_h \tag{1}$$

In depthwise convolution, each input channel is independently convolved with its corresponding kernel channel, thereby avoiding cross-channel computations. The computational complexity is given by Equation (2):

$$\text{FLOPs}_{\text{DW}} = W_o \times H_o \times C_i \times K_w \times K_h \tag{2}$$

Pointwise convolution is essentially a standard convolution operation with a kernel size of 1×1. In the forward propagation of the MBConv module, two pointwise convolutions and one depthwise convolution are sequentially performed. The ratio of computational complexity between this module (MBConv) and standard convolution is given by Equation (3):

$$\frac{\text{FLOPs}_{\text{MBConv}}}{\text{FLOPs}_{\text{Conv}}} = \frac{C_i \times C_o + C_i \times K_w \times K_h + C_i \times C_o}{C_o \times C_i \times K_w \times K_h} = \frac{K_w \times K_h + 2C_o}{C_o \times K_w \times K_h} \tag{3}$$

Assuming an output channel number of 256 and a convolution kernel size of 3×3, substituting these values into the ratio formula yields that the computational complexity of the convolution operations within the MBConv module is only 22.6% of that required by standard convolution.

The Effective Squeeze-and-Excitation (ESE) attention mechanism is a lightweight, improved module designed for channel-wise feature recalibration tasks. To address the channel dimensional collapse issue caused by the two fully connected layers (dimensionality reduction–expansion) paradigm employed by the classical SENet [24], the ESE module innovatively reconstructs the channel dependency modeling process. By directly learning global correlations among channels through a single fully connected layer and eliminating redundant dimensionality reduction operations, the ESE module maintains the integrity of original channel information while significantly reducing computational overhead. This design achieves accurate modeling of cross-channel interactions under the premise of reduced parameters. It provides an efficient computational foundation for decoupling fire features from complex backgrounds. In this paper, we propose a lightweight attention enhancement strategy, replacing the standard SE channel attention within the MBConv module with the ESE module, thereby constructing a composite convolutional unit termed EMB. The EMB module is integrated into the C2f architecture, replacing the original standard convolutional unit CBS (Conv, BN, SiLU), thus forming a novel C2f-EMB module.

In summary, the proposed multi-scale lightweight collaborative architecture effectively integrates the benefits of multi-scale network structure optimization and lightweight convolutional reconstruction of the C2f module. Specifically, by introducing a deeper $64\times$ down-sampling pathway and complementary feature fusion operations, the redesigned network significantly enhances the detection capability for large-scale targets encountered in construction site fire scenarios. Concurrently, the innovative lightweight C2f-EMB module, which leverages MBConv with the optimized ESE attention mechanism, addresses the increased computational overhead resulting from deeper network structures. This integrated approach balances semantic representation, feature extraction efficiency, and computational resource management, ultimately achieving superior performance suitable for practical applications in complex construction site environments. In the subsequent sections of this paper, we refer to this architecture as the “MSCE” structure.

3.2. Inverted Residual Spatial–Channel Attention Module

The complexity and variability of construction site environments make it challenging for the backbone network to effectively distinguish foreground targets from background noise during feature extraction. This frequently leads the model to mistakenly identify irrelevant background features as important information, thus reducing the accuracy of object detection. Reducing interference from background noise during feature extraction has become a critical challenge for enhancing detection performance. To address this issue, this paper proposes a novel spatial–channel attention mechanism termed iCBAM. The network architecture of this module is illustrated in Figure 6.

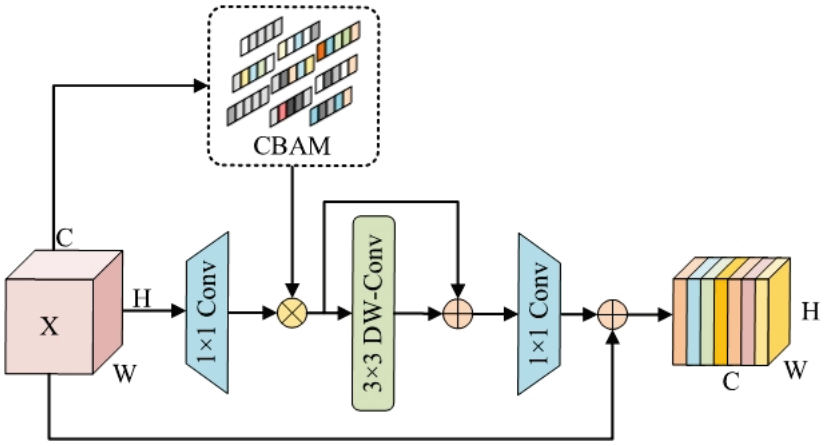


Figure 6. iCBAM attention mechanism network architecture.

This module integrates the channel–spatial collaborative modeling capability of the CBAM [25] attention mechanism with the efficient feature-propagation pathway of the inverted residual block [26] (iRB), significantly enhancing the network's feature representation for flame and smoke regions while maintaining controlled computational overhead. The input feature map first passes through the CBAM module, where channel and spatial attention weights are computed sequentially. The computed attention weights are multiplied element-wise with the output of the pointwise convolution layer in the iRB module, achieving joint calibration across both channel and spatial dimensions to emphasize fire-related features. Subsequently, depthwise separable convolutions are applied to further extract feature information. Finally, a pointwise convolution adjusts the output channel count to match that of the input, and a residual connection merges the original input features to mitigate the vanishing-gradient issue.

In this work, the attention module is embedded within a lightweight inverted residual structure to fully leverage its efficient information flow and gradient propagation advantages. iCBAM not only enhances the network's perception of critical regions while preserving the integrity of feature representations, but also effectively mitigates erroneous dependencies arising from excessive semantic focusing in deep-feature modeling by traditional attention mechanisms. This design suppresses the tendency of attention weights to overfit at high semantic levels, preventing the network from erroneously focusing on redundant regions unrelated to fire features, thereby enhancing the model's discriminative robustness in complex scenarios.

3.3. Improvement of the Boundary Loss Function

In construction site fire detection tasks, background occlusions and overlapping targets in complex construction environments significantly interfere with the accurate localization of candidate bounding boxes. The YOLOv8 model employs a combination of Distribu-

tion Focal Loss (DFL) and Complete Intersection over Union [27] (CIoU) loss to jointly optimize bounding box regression. The core formulas for CIoU are defined as follows in Equations (4)–(6):

$$CIoU = IoU - \frac{\rho\left(b, b_{gt}\right)^2}{c^2} - \alpha v \tag{4}$$

$$\alpha = \frac{v}{1 - IoU + v} \tag{5}$$

$$v = \frac{4}{\pi^2}\left(\arctan\frac{w_{gt}}{h_{gt}} - \arctan\frac{w}{h}\right)^2 \tag{6}$$

In the equations above, $\rho\left(b, b_{gt}\right)$ represents the Euclidean distance between the center points of the predicted bounding box and the ground-truth bounding box; α denotes the aspect ratio weighting factor; c is the diagonal length of the smallest enclosing box; and v is a measure of the aspect ratio distortion. Due to variations in data acquisition devices, shooting distances, and illumination conditions, the quality of collected fire images varies considerably, consequently leading to differences in the quality of annotated samples. The geometric characteristics of construction site fire images exhibit significant difference, and the original CIoU function neglects differences in learning difficulty among samples, making it prone to harmful gradient gains from low-quality samples (e.g., blurred flames or distorted smoke). For the above problems, this paper adopts the dynamic gradient optimization-based WIoUv3 [28] loss function, which balances the gradient contributions from samples of varying quality through an outlier degree mechanism. The formulas are defined as Equations (7)–(9):

$$L_{WIoU_{v1}} = R_{WIoU} * L_{IoU}, L_{IoU} = 1 - \frac{W_i H_i}{S_u} \tag{7}$$

$$R_{WIoU} = exp\left(\frac{\left(x - x_{gt}\right)^2 + \left(y - y_{gt}\right)^2}{W_g^2 + H_g^2}\right) \tag{8}$$

$$L_{WIoUv3} = r \cdot L_{WIoUv1}, \quad r = \frac{\beta}{\delta\alpha^{(\beta-\delta)}}, \beta = \frac{L_{IoU}^*}{\overline{L_{IoU}}} \tag{9}$$

Equation (8) represents the formulation of WIoU v1, which incorporates a distance attention mechanism to amplify the loss contribution of anchors with average quality while reducing the loss from high-quality anchors. WIoU v3 builds upon the distance attention mechanism introduced in WIoU v1 by incorporating a non-monotonic focusing factor r and an outlier degree β, where β is defined as the ratio between the exponential moving average and the simple average of the IoU. This metric reflects the quality of the anchor box and enables the model to adaptively adjust the gradient contributions of different samples. WIoU v3 assigns smaller gradient gains to samples exhibiting extreme outlier degrees, enabling the loss function to focus primarily on anchors of moderate quality. This strategy effectively mitigates harmful gradients caused by low-quality samples, thereby improving the detection performance of the enhanced model.

4. Experiments and Results

4.1. Evaluation Metrics

Evaluation metrics for fire detection systems primarily focus on accuracy and response speed. Therefore, this study employs Precision, Recall, mean Average Precision (mAP), and Giga Floating Point Operations Per second (GFLOPs) as the main evaluation metrics [29]. Firstly, it is essential to clarify several fundamental cases within detection results: True

Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). TP refers to the number of positive samples correctly predicted as positive; FP denotes the number of negative samples incorrectly classified as positive; FN represents the number of positive samples incorrectly classified as negative.

Precision represents the proportion of correctly predicted results among all predicted outcomes, reflecting the accuracy of the model's predictions. It is calculated as Equation (10):

$$\text{Precision} = \frac{TP}{TP + FP} \tag{10}$$

Recall represents the proportion of correctly predicted results among all actual positive instances, reflecting the comprehensiveness of the model's predictions. It is calculated as Equation (11):

$$\text{Recall} = \frac{TP}{TP + FN} \tag{11}$$

By setting precision as the Y-axis and recall as the X-axis, a rectangular coordinate system is established to plot the Precision-Recall (PR) curve. The area enclosed by the PR curve and the coordinate axes represents the Average Precision (AP). When the dataset contains multiple classes, the mean Average Precision (mAP) is the average of AP values across all categories, calculated as Equation (12):

$$\text{mAP} = \frac{1}{N}\sum_{i=1}^{N} \text{AP}_i \tag{12}$$

GFLOPs is a critical metric for measuring the computational complexity of a model, representing billions of floating-point operations executed per second. It quantifies the computational overhead of a model during inference by counting and normalizing the multiply–add operations within each computational module of the deep neural network into billions of floating-point operations.

4.2. Experimental Environment and Hyperparameter Settings

To ensure the scientific rigor and fairness of the experiments, all comparative experiments were conducted on the same software and hardware platform, with training hyperparameters strictly controlled. The experimental environment configuration is presented in Table 1:

Table 1. Configuration of the experimental environment and hyperparameters.

Category	Configuration	Value
Hardware Platform	CPU	Intel Xeon E5-2609 v4 @1.70 GHz
	GPU	Nvidia RTX 3090(24 GB) × 2
	Memory	Kingston DDR4 3200 Hz 128 GB
	Hard Disk	DELL SATA 32 TB
Software Platform	OS	Ubuntu 22.04 LTS
	Framework	PyTorch 2.2 CUDA 11.3
	Container	Docker v5.1.3
Hyperparameters	Epochs	300
	Patience	50
	Batch	16
	Learning Rate	0.01
	Optimizer	SGD
	Momentum	0.937

Among them, "Epochs" refers to the number of complete training iterations; "Patience" indicates the maximum number of consecutive epochs allowed without performance improvement before early stopping is triggered; "Batch" denotes the batch size, that is, the number of samples used for each parameter update; and "lr0" represents the initial learning rate set at the start of training.

4.3. Ablation Study

4.3.1. Results and Analysis

To validate the performance of the proposed algorithm for fire detection in complex construction site scenarios, systematic experiments were conducted on our self-built Fire-5000 dataset. The dataset was split into training and validation sets at a 4:1 ratio, ensuring a balanced class distribution. Experiments were conducted using YOLOv8s as the baseline model, ensuring a fair comparison under identical hyperparameter configurations.

To comprehensively assess the impact of each module on model performance, ablation experiments were conducted by incrementally introducing the proposed improvement strategies into the baseline YOLOv8s model. A multidimensional quantitative analysis of each strategy's impact on model performance was performed, and the results are presented in Table 2.

Table 2. Results of the ablation study.

ID	A	B	C	D	Precision	Recall	mAP	mAP50-95	Parame	GFLOPs	FPS
0					68.3	64.3	69.7	35.7	11126358	28.4	230
1	√				71.5	65.2	70.4	36.4	21158870	28.7	117
2		√			70.8	66.1	71.2	36.3	12444534	23.4	153
3			√		71.4	64.9	70	35.8	11428024	28.7	218
4				√	68.7	65.1	70.1	35.6	11126358	28.4	229
5	√	√			71.4	66.3	70.8	36.3	15217976	23.7	145
6	√	√	√		72.8	65.0	71.1	36.6	20519642	23.8	145
7	√	√	√	√	**72.9**	**67.3**	**71.3**	**37.2**	20519642	23.8	144

A refers to multi-scale detection; B refers to the C2f-EMB module; C denotes the iCBAM attention mechanism; and D refers to the WIoU loss function.

By incorporating an additional detection branch with a 64× down-sampling ratio into the baseline model, we aimed to improve its perceptual capability to large-scale fire targets. Experiment (1) demonstrates that this architecture increases the model's precision by 3.2% and raises mAP from 69.7% to 70.4%, thereby validating the effectiveness of the added scale branch for detecting large-scale targets. However, this modification also introduces additional convolutional pathways, increasing the parameter count to 21.15 million and reducing inference speed to 117 FPS, thereby significantly undermining the model's real-time detection performance.

In Experiment (2), a lightweight convolution-reconstructed C2f-EMB module was introduced into the original YOLOv8 architecture to replace the conventional C2f module. This module integrates the advantages of depthwise separable convolutions and bottleneck architectures, markedly reducing computational redundancy while enhancing feature propagation efficiency and inter-channel representation capacity, thereby effectively improving the model's discrimination performance for fire targets. Experimental results show that this improvement elevated the mAP to 71.2% and the recall to 66.1%, achieving the best performance among all single-module enhancement schemes and fully validating the practical value of the C2f-EMB structure in object-detection tasks under complex environments.

Given the stringent real-time performance requirements of fire-detection tasks, this study integrates the iCBAM attention mechanism at a single location above the SPPF module in the backbone network. Results from Experiment (3) demonstrate that this module

significantly improves the model's precision, yielding a 3.1% increase, but gains in other performance metrics are comparatively modest. Without compromising recall, iCBAM effectively enhances the model's discrimination of positive-class targets and suppresses interference from background noise.

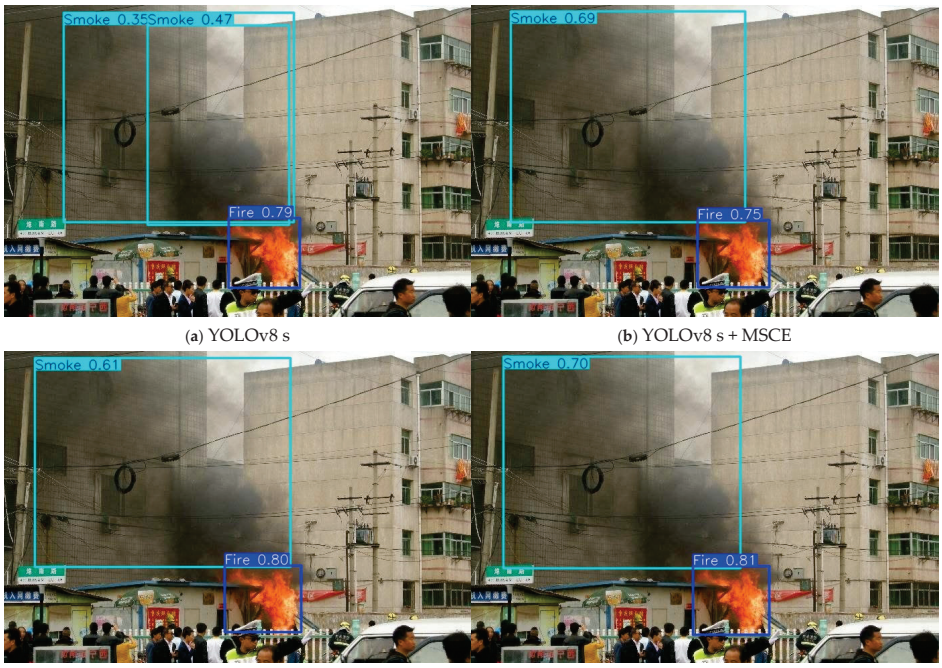
The improvement of the loss function primarily targets gradient allocation without altering the model's feature extraction and representation capabilities; therefore, it is difficult to yield a substantive performance leap. Results show that incorporating the WIoU loss function, without increasing the model's computational overhead, yields a little improvement in detection performance.

By jointly integrating the multi-scale detection structure and the lightweight feature-enhancement module C2f-EMB, we construct the proposed "Multi-Scale Lightweight Collaborative Architecture." Compared to the multi-scale structure, this architecture significantly enhances the model's recall (improving it by 1.1%) and increases mAP by 0.4%, while maintaining precision and mAP 50–90 performance. Furthermore, owing to the optimized design of the lightweight convolutional units, this scheme demonstrates superior computational efficiency, reducing the parameter count by 28% and the FLOPs by 17%, thereby achieving a favorable balance between detection performance and computational overhead. Experimental results indicate that, although the proposed architecture does not universally outperform every single optimization strategy on certain evaluation metrics, it achieves better comprehensive performance on key indicators such as precision, recall, mAP, and computational overhead. Finally, building upon the aforementioned optimizations, the iCBAM attention mechanism and the WIoU loss function are sequentially introduced to construct the final improved algorithm, YOLO-Fire, yielding a second performance leap. Compared to the baseline YOLOv8s model, YOLO-Fire achieves significant improvements across multiple core metrics: precision increases by 4.6%, recall by 3.0%, mAP by 1.6%, and mAP 50–90 by 1.5%, thereby fully demonstrating the superiority of the proposed algorithm.

4.3.2. Visualization Results

In the following sections, this paper will comparatively analyze the performance of the different enhancement strategies in practical detection scenarios. This comparative experiment involves the baseline model, YOLOv8s, the enhanced models, YOLOv8 + MSCE and YOLOv8 + MSCE + iCBAM, and the final model, YOLO-Fire. The aforementioned models were applied to detect the same construction site fire image, and the visualized comparative experimental results are presented below.

Figure 7 indicates that because the baseline YOLOv8 lacks an effective detection mechanism for large-scale objects, it erroneously fragments a single large smoke target into two independent instances, each assigned a low confidence score. With the incorporation of the MSCE architecture, the addition of a dedicated detection layer for large-scale objects equips the model with enhanced perception of wide-ranging targets, enabling accurate identification of smoke instances; however, the confidence scores still leave room for improvement. Upon incorporation of the iCBAM attention mechanism, the model's feature representation is strengthened, thereby facilitating the precise identification of smoke instances. Finally, by integrating multiple improvement strategies, the YOLOv8-Fire model is formed. This model combines the MSCE architecture's capability for effective large-scale object capture with the iCBAM attention mechanism's enhanced feature representation, thereby accurately identifying large-scale smoke targets while significantly improving the confidence scores of detected objects. The aforementioned visualization experiments validate the effectiveness of the proposed fusion algorithm in practical fire-detection scenarios.



(**a**) YOLOv8 s (**b**) YOLOv8 s + MSCE

(**c**) YOLOv8 s +MSCE + iCBAM (**d**) YOLO-Fire

Figure 7. Comparison of the practical effectiveness of the optimization algorithm.

To intuitively illustrate the optimization effects of the YOLO-Fire model, this section plots the curves of key performance metrics over training iterations. A comparative analysis with the baseline YOLOv8s model clearly demonstrates the enhancements in detection performance afforded by the proposed improvement strategies.

As shown in Figure 8, the YOLO-Fire model attains its peak performance at the 150th iteration. In the latter half of the training curve, all four evaluation metrics of YOLO-Fire significantly exceed those of the baseline model, indicating that the proposed algorithm is better suited to fire-detection tasks in construction site scenarios.

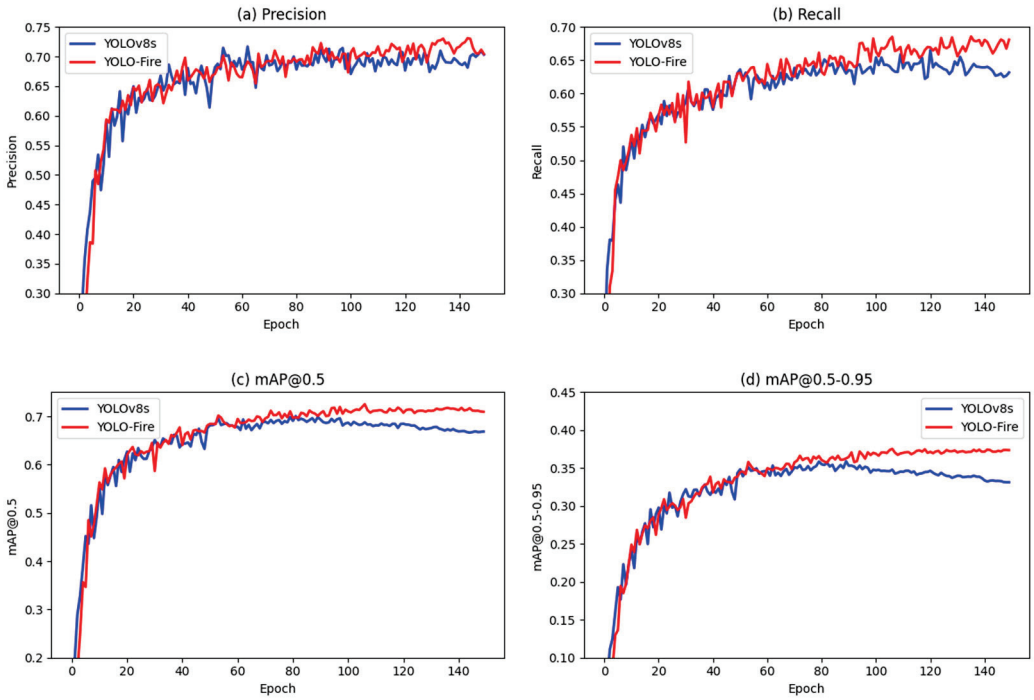


Figure 8. Iteration curves of performance metrics.

4.4. Comparative Experiments with Classical/SOTA Models

To further validate the effectiveness of the proposed algorithm, this section conducts a comprehensive comparative analysis of its performance against classical and state-of-the-art (SOTA) algorithms on our self-built construction site fire dataset, thereby more clearly demonstrating the advantages and characteristics of the proposed method. Since the different models differ in size specifications and underlying principles, comparing computational complexity and real-time metrics such as FPS would lack rigor. The detection performance of the selected models on the self-built construction site fire dataset was compared across four metrics: precision, recall, mAP, and mAP50-90.

Table 3. Performance comparison of various models.

Model	Precision (%)	Recall (%)	mAP (%)	mAP 50-95 (%)
YOLOv8	68.3	64.3	69.7	35.7
YOLOv5	71.2	63.2	70.1	35.2
YOLOv9 [30]	71.7	65.6	70.9	36.5
YOLOv10	69	36.6	69.3	35.6
YOLOv11	69.9	64.1	70.4	36.2
RT-DETR	72.1	61.2	65.1	31.8
Hyper-YOLO [31]	68.1	65.3	68.9	35.4
Mamba-YOLO [32]	70.3	65.9	71	**37.3**
YOLO-Fire	**72.9**	**67.3**	**71.3**	37.2

Analysis of the data in Table 3 reveals that the YOLO-Fire algorithm proposed in this paper exhibits significant performance advantages in construction site fire detection tasks. In terms of precision, the YOLO-Fire algorithm achieved 72.9%, which is not only markedly higher than the baseline YOLOv8s model 68.3% but also exceeds the current state-of-the-art RT-DETR [33] model 72.1%. With respect to recall, YOLO-Fire likewise demonstrates outstanding performance, achieving 67.3% and comprehensively outperforming all other compared models. Furthermore, on the mAP metric reflecting overall detection performance, YOLO-Fire achieved a best result of 71.3%; on the more stringent mAP50–90 metric, YOLO-Fire reached 37.2%, closely approaching the top-performing Mamba-YOLO 37.3%. In summary, the YOLO-Fire algorithm demonstrates remarkable performance advantages and high stability in construction site fire detection tasks, thereby fully validating the effectiveness of the proposed MSCE architecture and attention mechanism optimization strategies.

4.5. Cross-Domain Generalization Evaluation

To further evaluate the generalization capability of the proposed YOLO-Fire algorithm across diverse fire scenarios and to validate its effectiveness and robustness in a complex application environment. In this subsection, we select the publicly available D-Fire and FASDD [34] fire-detection datasets to conduct cross-domain generalization performance comparison experiments. These two datasets differ markedly from the construction site fire dataset used in this study in terms of fire scenario types, background complexity, and object-scale distributions, thus providing an objective assessment of the model's generalization and adaptability to unknown environments.

4.5.1. FASDD Dataset

In this study, we employ the FASDD dataset—a remote sensing–based flame and smoke target detection dataset for deep learning—publicly released by Wuhan University in 2024 via the Scientific Data Bank. The dataset comprises fire images captured at varying

viewing distances, across different scenarios and illumination conditions, and acquired using a variety of visual sensors. The dataset comprises on the order of 10^5 images, containing 113,154 flame object instances and 73,072 smoke object instances, making it the largest and most broadly representative flame-and-smoke detection dataset to date. The included scenarios are illustrated in Figure 9.



Figure 9. Fire images from the FASDD dataset.

This experiment compared the detection performance of the baseline YOLOv8s model and the proposed YOLO-Fire on the FASDD dataset; the results are shown in Table 4.

Table 4. Comparative experiments on the FASDD dataset.

Target Objects	YOLOv8s			YOLO-Fire		
	Precision	**Recall**	**mAP**	**Precision**	**Recall**	**mAP**
Fire	71.5	64.7	72.6	71.7	65.7	73.4
Smoke	81.9	72.7	81.6	84.5	72.2	82.4
Average	76.7	68.7	77.1	78.1	69	77.9

Compared with the baseline model, the proposed YOLO-Fire algorithm achieved improvements of 1.3%, 0.3%, and 0.8% in precision, recall, and mAP on the FASDD dataset, respectively, thereby validating the effectiveness of the proposed enhancement strategies in cross-domain scenarios. However, the recall for the smoke category exhibited a slight decline of 0.5 percentage points. Analysis indicates that the primary reason is that, in the FASDD dataset, the average scale of smoke targets captured from UAV and satellite perspectives is smaller compared to those in our self-built dataset. This results in an excessively large receptive field in the large-object detection layer during feature fusion, which dilutes local fine-grained information and causes missed detections of small-scale smoke targets. Nevertheless, YOLO-Fire still outperforms the baseline model across all other key metrics, indicating that the multi-scale collaborative optimization architecture (MSCE) and the dynamic gradient allocation strategy proposed in this paper substantially enhance the model's feature decoupling capability, thereby achieving stable generalization across diverse detection tasks.

4.5.2. D-Fire Dataset

The D-Fire [35] (Drone-based Fire) dataset is a UAV fire-imagery dataset specifically designed by the Brazilian GAIA research team for machine learning and object-detection algorithms, comprising over 21,000 images. The dataset's images are categorized into four classes: 1164 images containing only flames, 5867 images containing only smoke, 4658 images containing both flames and smoke, and 9838 distractor images containing neither flames nor smoke. A total of 14,692 flame instances and 11,865 smoke instances were annotated [36]. The D-Fire dataset was collected primarily in outdoor wildfire environments

such as forests and grasslands, which differ substantially from the construction site fire scenarios addressed in this study. Therefore, selecting the D-Fire dataset for cross-domain generalization experiments facilitates an objective validation of the proposed algorithm's generalization capability and adaptability across different fire scenarios, thereby ensuring scientific rigor. The models compared in this experiment are YOLOv8, YOLOv10 [37], YOLOv11, and the proposed YOLO-Fire; the experimental results are presented in Table 5.

Table 5. Performance on the D-Fire dataset.

Model	Precision	Recall	AP		mAP
			Fire	Smoke	
YOLOv8	78.3	72.2	72.6	84.4	78.5
YOLOv10	79.9	72.4	72.8	85.5	79.2
YOLOv11	79.1	73.3	73.2	85	79.1
YOLO-Fire	82.3	73.7	74.4	85.9	80.1

As shown by the comparative experiments on the D-Fire dataset (Table 5), the YOLO-Fire algorithm proposed in this paper demonstrates clear performance advantages. The YOLO-Fire algorithm outperforms all comparison models across precision, recall, and mAP metrics. The experimental results indicate that the proposed optimization strategies effectively enhance the model's adaptability and generalization performance across diverse fire scenarios, enabling the YOLO-Fire algorithm to maintain stable and superior performance in complex and variable fire-detection environments, thus offering clear application advantages.

Figure 10 illustrates the visualized performance of YOLO-Fire and other one-stage object detection algorithms on the D-Fire dataset, highlighting the differences in detection performance under real-world fire scenarios.

Figure 10a presents the detection results of each algorithm under environmental interference conditions. In this instance, YOLOv8 missed a fire instance due to occlusion by a foreground vehicle, whereas YOLOv10 erroneously identified the worker's red jacket on the right as a fire target. In contrast, both YOLOv11 [38] and the proposed YOLO-Fire algorithm successfully detected all fire targets, with YOLO-Fire achieving higher confidence scores.

Figure 10b presents the detection results in a multi-scale wildfire scenario. YOLOv8 failed to detect one small-scale flame instance, whereas the other algorithms successfully identified all targets, demonstrating superior multi-scale adaptability and detection precision.

Figure 10c illustrates detection performance for targets at the image periphery. Among all compared methods, only the proposed YOLO-Fire algorithm accurately detected the flame instance located at the left border of the image and achieved the highest confidence score.

In summary, both in theoretical metrics and practical detection performance, YOLO-Fire significantly outperforms comparable algorithms, indicating that the proposed method possesses strong cross-domain generalization capabilities.



Figure 10. Analysis of the generalization performance of YOLO-Fire. (**a**) Environmental interference scenarios. (**b**) Small-scale fire target. (**c**) Edge target detection.

5. Discussion

Although the YOLO-Fire algorithm has achieved significant improvements in detecting fires within construction site environments, several limitations remain that deserve further attention. Firstly, while the algorithm demonstrates adequate real-time performance under current hardware conditions (achieving 144 FPS), the computational complexity of the proposed model could still pose deployment challenges in scenarios with extremely limited computational resources or embedded devices [39]. Future work should explore additional lightweight optimization techniques to further reduce the computational overhead.

Secondly, despite robust performance in typical fire scenarios, the proposed model exhibits slight weaknesses when dealing with extremely small-scale smoke targets, as revealed in cross-domain experiments on the FASDD dataset. This indicates that the multi-scale structure could potentially dilute fine-grained local features during feature fusion, affecting the detection of very small targets. Subsequent studies should consider adaptive feature fusion mechanisms to better preserve and utilize fine-grained information.

Lastly, the current dataset, though comprehensive, still lacks sufficient representation of extremely rare fire scenarios, such as chemical or electrical fires specific to certain construction materials or methods. Expanding dataset coverage through targeted data collection could further enhance the robustness and practical applicability of the algorithm.

6. Conclusions

This study proposes an improved YOLOv8-based fire detection algorithm named YOLO-Fire, specifically optimized for construction site environments. Key innovations include a multi-scale lightweight collaborative architecture (MSCE) enhancing detection across diverse scales, the iCBAM attention module improving feature discrimination and suppressing background interference, and the dynamic gradient allocation strategy utilizing the WIoUv3 loss function. Experimental validation on the self-constructed Fire-5000 dataset demonstrates significant improvements: precision increased by 4.6%, recall by 3.0%, mAP by 1.6%, and mAP50-95 by 1.5% compared to baseline YOLOv8s. Furthermore, cross-domain tests on public datasets FASDD and D-Fire confirm the robustness and generalizability of YOLO-Fire. Overall, the proposed model effectively addresses the complexity of construction site fire detection tasks, achieving a balance between detection accuracy and computational efficiency.

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

Funding: This research was funded by Innovation Capacity Promotion Project for Technology-based SMEs in Shandong Province: "AI-empowered Smart Construction Site Platform Project", grant number 2023TSGC0877.

Institutional Review Board Statement: The study in this paper did not involve humans or animals.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

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

References

1. Alkhatib, A.A. A review on forest fire detection techniques. *Int. J. Distrib. Sens. Netw.* **2014**, *10*, 597368. [CrossRef]
2. Jin, C.; Wang, T.; Alhusaini, N.; Zhao, S.; Liu, H.; Xu, K.; Zhang, J. Video fire detection methods based on deep learning: Datasets, methods, and future directions. *Fire* **2023**, *6*, 315. [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. Choi, S.; Kim, S.; Jung, H. Optimized Faster R-CNN with Swintransformer for Robust Multi-Class Wildfire Detection. *Fire* **2025**, *8*, 180. [CrossRef]
5. Cheknane, M.; Bendouma, T.; Boudouh, S.S. Advancing fire detection: Two-stage deep learning with hybrid feature extraction using faster R-CNN approach. *Signal Image Video Process.* **2024**, *18*, 5503–5510. [CrossRef]
6. 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]
7. Mukhiddinov, M.; Abdusalomov, A.B.; Cho, J. A wildfire smoke detection system using unmanned aerial vehicle images based on the optimized YOLOv5. *Sensors* **2022**, *22*, 9384. [CrossRef] [PubMed]
8. Liu, H.; Zhang, F.; Xu, Y.; Wang, J.; Lu, H.; Wei, W.; Zhu, J. Tfnet: Transformer-based multi-scale feature fusion forest fire image detection network. *Fire* **2025**, *8*, 59. [CrossRef]

9. Li, Y.; Nie, L.; Zhou, F.; Liu, Y.; Fu, H.; Chen, N.; Dai, Q.; Wang, L. Improving Fire and Smoke Detection with You Only Look Once 11 and Multi-Scale Convolutional Attention. *Fire* **2025**, *8*, 165. [CrossRef]
10. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
11. Jocher, G.; Stoken, A.; Borovec, J.; Changyu, L.; Hogan, A.; Diaconu, L.; Poznanski, J.; Yu, L.; Rai, P.; Ferriday, R. *ultralytics/yolov5, version 3.0*; Zenodo: Geneva, Switzerland, 2020.
12. Sohan, M.; Sai Ram, T.; Rami Reddy, C.V. A review on yolov8 and its advancements. In Proceedings of the International Conference on Data Intelligence and Cognitive Informatics, Tirunelveli, India, 18–20 November 2024; pp. 529–545.
13. Wang, C.-Y.; Liao, H.-Y.M.; Yeh, I.-H. Designing network design strategies through gradient path analysis. *arXiv* **2022**, arXiv:2211.04800.
14. Ramachandran, P.; Zoph, B.; Le, Q.V. Searching for activation functions. *arXiv* **2017**, arXiv:1710.05941.
15. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
16. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake, UT, USA, 18–23 June 2018; pp. 8759–8768.
17. Bakurov, I.; Buzzelli, M.; Schettini, R.; Castelli, M.; Vanneschi, L. Structural similarity index (SSIM) revisited: A data-driven approach. *Expert Syst. Appl.* **2022**, *189*, 116087. [CrossRef]
18. Chen, G.; Zhou, H.; Li, Z.; Gao, Y.; Bai, D.; Xu, R.; Lin, H. Multi-scale forest fire recognition model based on improved YOLOv5s. *Forests* **2023**, *14*, 315. [CrossRef]
19. Rasheed, A.F.; Zarkoosh, M. Optimized YOLOv8 for multi-scale object detection. *J. Real-Time Image Process.* **2025**, *22*, 6. [CrossRef]
20. 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 Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
21. Zhou, N.; Gao, D.; Zhu, Z. YOLOv8n-SMMP: A Lightweight YOLO Forest Fire Detection Model. *Fire* **2025**, *8*, 183. [CrossRef]
22. Koonce, B. EfficientNet. In *Convolutional Neural Networks with Swift for Tensorflow: Image Recognition and Dataset Categorization*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 109–123.
23. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake, UT, USA, 18–23 June 2018; pp. 4510–4520.
24. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake, UT, USA, 18–23 June 2018; pp. 7132–7141.
25. 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.
26. Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; pp. 6105–6114.
27. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU loss: Faster and better learning for bounding box regression. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 12993–13000.
28. Tong, Z.; Chen, Y.; Xu, Z.; Yu, R. Wise-IoU: Bounding box regression loss with dynamic focusing mechanism. *arXiv* **2023**, arXiv:2301.10051.
29. Padilla, R.; Netto, S.L.; Da Silva, E.A. A survey on performance metrics for object-detection algorithms. In Proceedings of the 2020 International Conference on Systems, Signals and Image Processing (IWSSIP), Niterói, Brazil, 1–3 July 2020; pp. 237–242.
30. 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.
31. Feng, Y.; Huang, J.; Du, S.; Ying, S.; Yong, J.-H.; Li, Y.; Ding, G.; Ji, R.; Gao, Y. Hyper-yolo: When visual object detection meets hypergraph computation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *47*, 2388–2401. [CrossRef]
32. Wang, Z.; Li, C.; Xu, H.; Zhu, X. Mamba YOLO: SSMs-based YOLO for object detection. *arXiv* **2024**, arXiv:2406.05835.
33. Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; Chen, J. Detrs beat yolos on real-time object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 16965–16974.
34. Wang, M.; Jiang, L.; Yue, P.; Yu, D.; Tuo, T. FASDD: An open-access 100,000-level flame and smoke detection dataset for deep learning in fire detection. *Earth Syst. Sci. Data Discuss.* **2023**, *2023*, 1–26.
35. de Venâncio, P.V.A.; Campos, R.J.; Rezende, T.M.; Lisboa, A.C.; Barbosa, A.V. A hybrid method for fire detection based on spatial and temporal patterns. *Neural Comput. Appl.* **2023**, *35*, 9349–9361. [CrossRef]
36. de Venancio, P.V.A.; Lisboa, A.C.; Barbosa, A.V. An automatic fire detection system based on deep convolutional neural networks for low-power, resource-constrained devices. *Neural Comput. Appl.* **2022**, *34*, 15349–15368. [CrossRef]

37. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J. Yolov10: Real-time end-to-end object detection. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 107984–108011.
38. Khanam, R.; Hussain, M. Yolov11: An overview of the key architectural enhancements. *arXiv* **2024**, arXiv:2410.17725.
39. Zhao, C.; Zhao, L.; Zhang, K.; Ren, Y.; Chen, H.; Sheng, Y. Smoke and Fire-You Only Look Once: A Lightweight Deep Learning Model for Video Smoke and Flame Detection in Natural Scenes. *Fire* **2025**, *8*, 104. [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.

big data and cognitive computing

MDPI

Systematic Review

A Systematic Literature Review of Artificial Intelligence in Prehospital Emergency Care

Omar Elfahim [1,2,*], Kokou Laris Edjinedja [1], Johan Cossus [3], Mohamed Youssfi [2], Oussama Barakat [1] and Thibaut Desmettre [4]

[1] SINERGIES Laboratory, Université Marie et Louis Pasteur, 16 Route de Gray, 25000 Besançon, France; kokou_laris.edjinedja2@univ-fcomte.fr (K.L.E.); oussama.barakat@univ-fcomte.fr (O.B.)
[2] 2IACS Laboratory, École Normale Supérieure de l'Enseignement Technique de Mohammedia (Morocco) (ENSET), University Hassan II of Casablanca, Bd Hassan II, Mohammédia 28830, Morocco; med@youssfi.net
[3] Emergency Department (SAU-SMUR), Centre Hospitalier Universitaire (CHU), Université Marie et Louis Pasteur, 3 Bd Alexandre Fleming, 25030 Besançon, France; jcossus@chu-besancon.fr
[4] Emergency Department, HUG Genève, Rue Gabrielle-Perret-Gentil 4, 1205 Geneva, Switzerland; thibaut.desmettre@hcuge.ch
* Correspondence: omar.elfahim@univ-fcomte.fr

Abstract

Background: The emergency medical services (EMS) sector, as a complex system, presents substantial hurdles in providing excellent treatment while operating within limited resources, prompting greater adoption of artificial intelligence (AI) as a tool for improving operational efficiency. While AI models have proved beneficial in healthcare operations, there is limited explainability and interpretability, as well as a lack of data used in their application and technological advancement. *Methods:* The scoping review was conducted according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines for scoping reviews, using PubMed, IEEE Xplore, and Web of Science, with a procedure of double screening and extraction. The search included articles published from 2018 to the beginning of 2025. Studies were excluded if they did not explicitly identify an artificial intelligence (AI) component, lacked relevance to emergency department (ED) or prehospital contexts, failed to report measurable outcomes or evaluations, or did not exploit real-world data. We analyzed the data source used, clinical subclasses, AI domains, ML algorithms, their performance, as well as potential roles for large language models (LLMs) in future applications. *Results:* A comprehensive PRISMA-guided methodology was used to search academic databases, finding 1181 papers on prehospital emergency treatment from 2018 to 2025, with 65 articles identified after an extensive screening procedure. The results reveal a significant increase in AI publications. A notable technological advancement in the application of AI in EMS using different types of data was explored. *Conclusions:* These findings highlighted that AI and ML have emerged as revolutionary innovations with huge potential in the fields of healthcare and medicine. There are several promising AI interventions that can improve prehospital emergency care, particularly for out-of-hospital cardiac arrest and triage prioritization scenarios. *Implications for EMS Practice:* Integrating AI methods into prehospital care can optimize the use of available resources, as well as triage and dispatch efficiency. LLMs may have the potential to improve understanding and assist in decision-making under pressure in emergency situations by combining various forms of recorded data. However, there is a need to emphasize continued research and strong collaboration between AI experts and EMS physicians to ensure the safe, ethical, and effective integration of AI into EMS practice.

Keywords: literature review; prehospital emergency; artificial intelligence; deep learning; machine learning; PRISMA; large language models; emergency medical services; LLMs

1. Introduction

In recent years, the pressure on EMS providers has been increasing due to the rising workload at dispatch centers, largely driven by population growth and an aging demographic. At the same time, there has been an increase in the occurrence of global pandemics [1] and a growing frequency of natural disasters [2], such as floods, wildfires, and earthquakes. For example, emergency activities in France have increased by 42% in the past 10 years, some coming directly from emergency rooms and others relying on telephone regulations that are essential for establishing ambulance priorities and emergency event priorities. This surge in demand has highlighted the need for innovative solutions to support emergency personnel and improve the overall delivery of emergency medical services. In order to handle such scenarios, emergency response systems need to be modernized. This requires advanced techniques that use artificial intelligence (AI) to help manage prehospital operations and control emergency calls, allowing for quicker reaction times and more effective decision-making. In this work, we explore the potential of AI to address some of these challenges by streamlining operations, enhancing decision-making processes, and ultimately improving patient outcomes. The development of timely and effective emergency management (EM) processes has become more important, with the goal of improving the coordination and efficiency of various emergency service components while planning for and responding to urgent occurrences. The EM architecture is organized around a set of decision-making difficulties that are divided into three key stages—call reception and triage, dispatch and on-scene response, and post-event analysis and reporting. Each step has unique subproblems that must be solved in order to provide a smooth and efficient emergency response system. For example, in North America, AI technologies are increasingly being used to help EMS crews with emergency call triage [3], predictive analytics, and real-time decision assistance [4]. These developments are intended to improve response times, more efficiently manage resources, and deliver important information to EMS and drive professionals during emergencies. In contrast, the employment of AI in EMS is less common in France and several German regions. Cultural variations, legal contexts, and varied amounts of technical infrastructure all contribute to the sluggish adoption of AI in various regions.

Healthcare is one of the most promising application domains for ML and DL, and natural language processing (NLP) is combined with machine learning (ML) [5] and DL to facilitate the extraction of relevant information from diverse clinical data sources across various tasks. Deep learning (DL) is a subset of ML based on multi-layered artificial neural networks, which can be further utilized to solve complex problems. The application of deep learning in healthcare has had great success and amazing performance improvements [6], including medical imaging analysis (e.g., detection of diabetic retinopathy) [7,8], prediction of outpatient disease risk, and early detection of heart failure [9]. In addition, to improve the safety of medications, researchers have introduced a variety of approaches using deep learning to help with medication recommendations based on patient data, such as diagnoses, procedures, and medication history extracted from longitudinal EHRs [10].

Figure 1 provides an overview of the review along with the main questions being addressed. Additionally, it shows the sub-aspects of each of these questions. The following sections address these seven important questions and their sub-aspects. First, we summarize AI technologies used in EMS and emergency pathways from the initial call

to emergency department admission, as used in evaluations. Then, we emphasize the optional but critical stages required for the use of AI in real-world scenarios of emergencies. These include data challenges in prehospital care, such as data linking complexity, privacy concerns, and multimodal data vectorization, which offer significant obstacles to AI application, covering both big data architectures in EMS systems and different data modalities that have been employed and integrated in this field. Finally, we identify some recent LLM applications in this evolving field, emphasizing the key challenges and benefits of adopting LLMs in prehospital care to improve real-time emergency response.

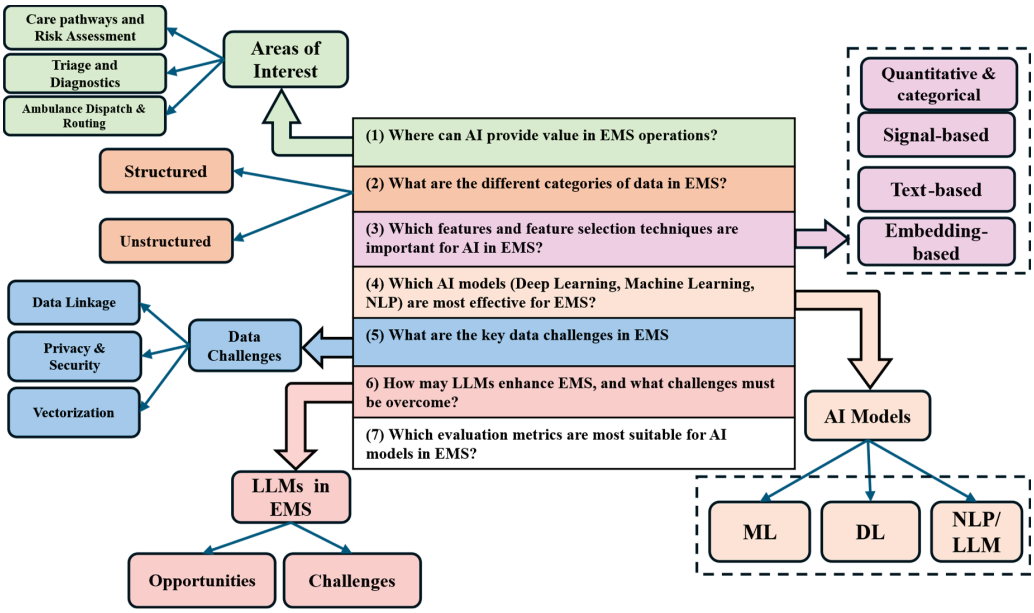


Figure 1. Survey overview.

This paper provides an overview of the current state of AI integration in EMS, focusing on successful applications, potential benefits, and existing challenges of AI in this field. By examining both the advancements and the obstacles faced by EMS providers globally, with a specific focus on the French context, we aim to highlight how AI can be used to improve prehospital emergency care and cope with the growing demands of EMS systems. This paper contributes significantly to the field of prehospital medical care by rigorously analyzing the current literature:

- It evaluates and organizes contributions to AI applications in prehospital emergency care, offering a complete overview of current developments.
- It investigates the constraints and limits connected with the use of AI in emergency medical care, highlighting important roadblocks that must be overcome.
- It recommends future research directions, indicating areas that require more inquiry to progress the discipline and overcome present limits.
- This essay also examines future prospects for large language models (LLMs) and generative AI in prehospital emergency care, emphasizing their potential to revolutionize and improve EMS processes.

The rest of this paper is arranged as follows: Section 2, based on the methodology, defines the research topics, search strategy, selection criteria, selection techniques, and study quality evaluation. In addition, it examines both included and excluded papers, creating a classification of AI applications in emergency medical care. Section 3 highlights the benefits of AI in emergency medical care, including predictive modeling, resource allocation, triage, diagnostics, and the nascent function of large language models (LLMs) in call regulation, while also discussing ethical issues and workflow integration. Section 4

addresses the major issues faced in the examined studies and proposes prospective areas for future research.Section 5 discusses database integration and data-centric methodologies, including big data architectures, data acquisition, storage, processing, and analytics, along with concerns about privacy, linkage, integrity, evaluation metrics, explainability, and data quality. Finally, Section 6 completes the work by summarizing our findings and suggesting future research topics.

2. Materials and Methods

2.1. An Overview of Emergency Call Center Operations

The Call Reception and Dispatch Center (CRRA 15) manages EMS call operations in France. It operates around the clock to address medical emergencies reported through various emergency numbers, including the SAMU (15), the Police (17), the Fire Brigade (18), and the pan-European emergency number (112). Specialized staff prioritize each call and, if required, deploy an emergency vehicle accompanied by a competent medical team. Figure 2 shows a detailed flow of the call screening procedure in the CRRA centers.

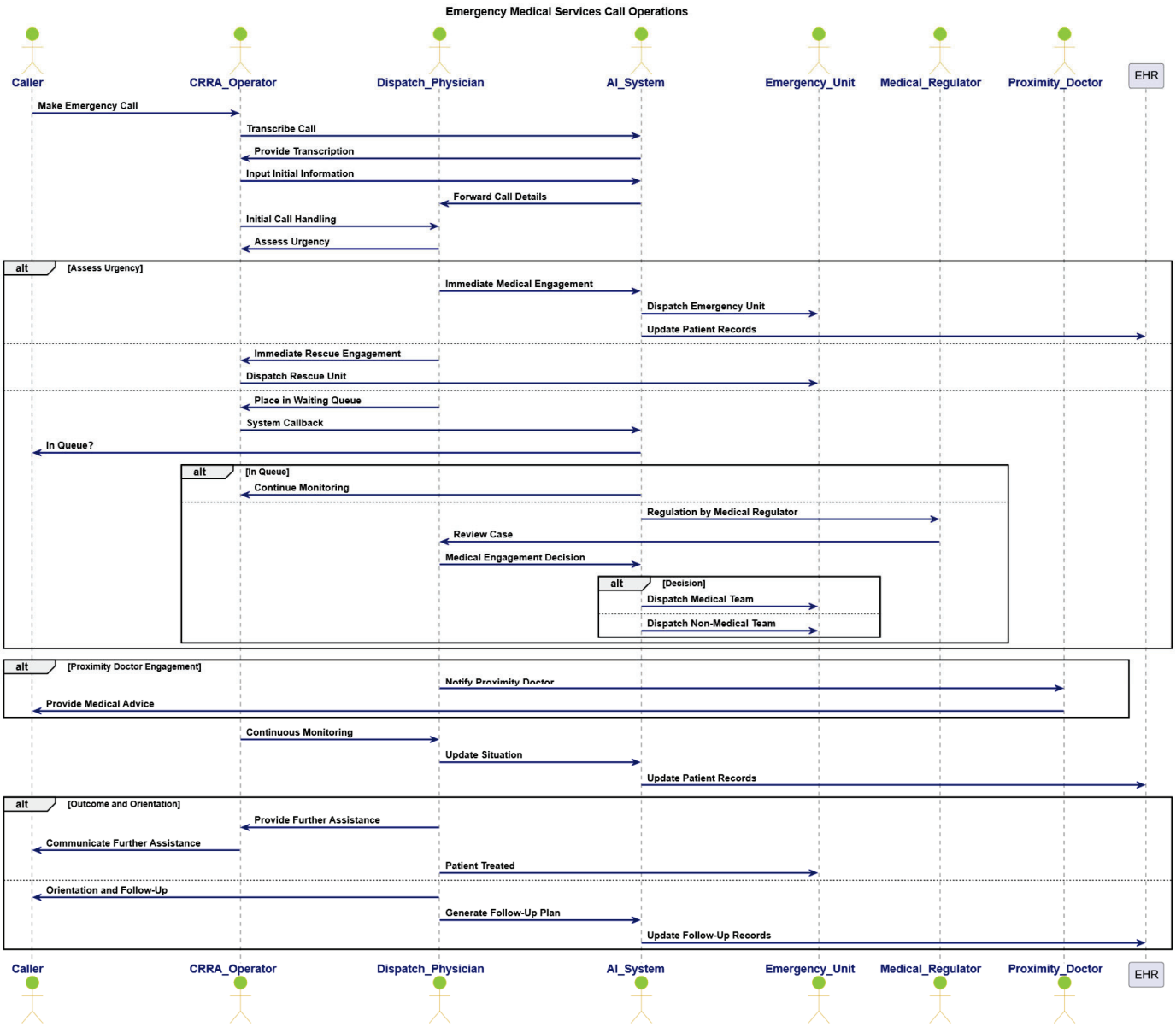


Figure 2. Emergency pathway: from initial call to emergency department admission.

Calls answered at the dispatch centers of the French EMS system come from a variety of sources. SAMU (Service d'Aide Médicale Urgente) is a dedicated line for medical emergencies, offering counseling and referral services in cases of sickness and medicine. In the event of a medical emergency, calls are handled and managed by CRRA 15.

Numbers 17 and 18 are dedicated lines for police and fire emergencies, respectively. These calls are evaluated for the nature of the emergency and, if they are health-related, transferred to EMS dispatch centers; otherwise, they are routed to the relevant authorities. The medical advice number (3966) is a number that offers medical advice and information. Calls about medical crises are routed to EMS dispatch centers and treated as emergency calls. Regardless of the source, every call is treated in the same manner. The premedical emergency staff conducts an initial screening to determine the nature of the emergency, whether it is an ongoing case (routine call), a new case (emergency call), or a false alarm. Routine calls and false alarms are logged and concluded, while new emergencies are prioritized and may lead to the delivery of an emergency vehicle or further medical assistance if the situation is less critical.

2.2. An Overview of AI

Artificial intelligence (AI) is increasingly recognized as a transformative force in EMS, particularly in the area of prehospital care. AI technologies such as machine learning, natural language processing, and deep learning are capable of extracting actionable insights from large and heterogeneous patient data sources, providing emergency care professionals with significant assistance (Figure 3). The use cases of AI in EMS have the potential to revolutionize the way in which EMS is delivered, improving patient outcomes, reducing costs, and increasing access to care. The integration of AI into EMS workflows holds promise for enhancing diagnostic precision, forecasting patient trajectories, supporting clinical decision-making, optimizing dispatch coordination, streamlining triage procedures, and facilitating the development of customized treatment plans. Additionally, AI-powered tools can help emergency service decision makers identify early warning signs of diseases and conditions [11], enabling early intervention and prevention [12]. This can significantly improve the efficacy and efficiency of emergency response services, which eventually improves patient outcomes. Simultaneously, the development of Big Data analytics has made a substantial contribution to the effective application of AI in healthcare environments. Sophisticated AI models are being used to find hidden clinical trends that would be hard to find otherwise, as healthcare systems produce datasets that are more complicated. Clinical decisions that are better informed and supported by data are made possible by these capabilities. ML models for structured datasets, including support vector machines and neural networks, as well as DL and NLP techniques for handling unstructured data, are examples of commonly used AI techniques. Large language models (LLMs) have lately shown promise in synthesizing clinical narratives and assisting in real-time decision-making.

Machine learning (ML) approaches, particularly supervised methods such as support vector machines (SVMs), random forests, and logistic regression, have been used for classification and regression tasks such as automating the classification of emergency call types, estimating clinical urgency, and improving dispatcher decisions. In addition, unsupervised techniques such as principal component analysis (PCA) and clustering algorithms demonstrated potential when it comes to uncovering latent structures in EMS data, such as detecting temporal clusters of high-risk occurrences or mapping geographic zones with higher call density. Furthermore, boosting and bagging are examples of ensemble techniques that further increase prediction performance. Deep learning (DL) provides improved processing capabilities for complicated large-scale EMS datasets. The potential of convolutional neural networks (CNNs) in image-based decision assistance is being

investigated, including determining the severity of injuries from field-acquired visuals. Applications such as forecasting demand trends and monitoring patient vitals during prehospital care have benefited from the use of sequential models such as recurrent neural networks (RNN), long short-term memory (LSTM), and gated recurrent units (GRU) on time-dependent data. Furthermore, autoencoders and Transformer-based architectures are also becoming more popular for applications like automatic incident report summarizing and anomaly identification in dispatcher–caller communication patterns. Improved situational awareness and quicker clinical response are two benefits of these technologies, especially when combined with real-time EMS operations.

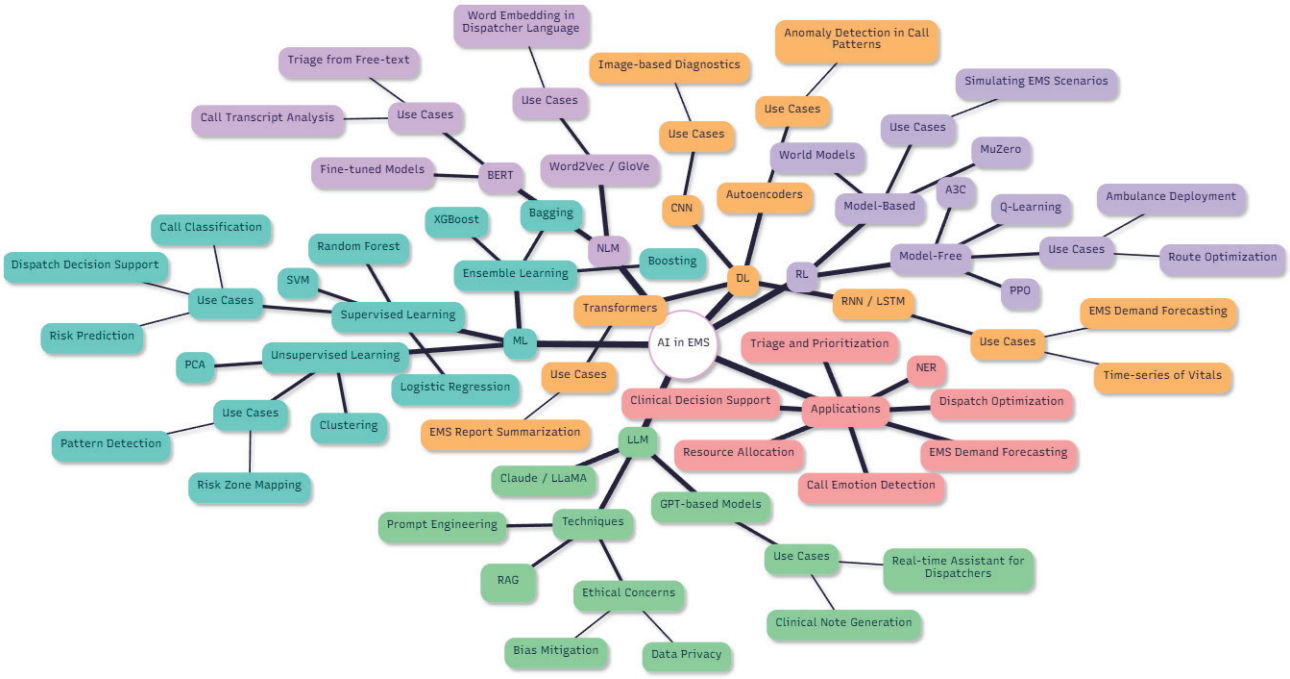


Figure 3. Mind map summary of the AI domains—machine learning, deep learning, natural language models, large language models, and reinforcement learning—along with representative algorithms and particular uses in prehospital emergency care, such as resource management, emotion recognition, dispatch optimization, and triage.

Natural language processing (NLP) and large language models (LLMs) have expanded the scope of AI in EMS communication analysis. Early NLP techniques relied heavily on frequency-based models; for example, a "bag-of-words" model makes decisions based on their relative frequency in the text, ignoring their order [13], "TF-IDF" introduced term weighting to emphasize informative words across documents. However, both approaches lacked semantic knowledge. Significant advancements in NLP have been made over the last few years through the use of DL, by introducing embedding models like Word2Vec and GloVe [14]. These models are pretrained on extensive datasets to capture semantic and syntactic patterns and mainly use the "Transformer" architecture to handle sequential data effectively. The concept of scaling laws is crucial to their performance, as it highlights how increasing both model size and data can lead to significant gains in understanding and generating human-like language and contextual knowledge applicable across domains. More recently, LLMs such as GPT-4 [15] and LLaMA [16] have been adapted for use in virtual agents designed to support clinical documentation, dispatcher training, and real-time question-answering. Techniques such as prompt engineering, agentic AI [17], and retrieval-augmented generation (RAG) improve the contextual precision of these models. However, issues are related to data privacy and other ethical considerations, and integration into EMS workflows.

Reinforcement learning (RL) has emerged as a powerful paradigm for optimizing real-time EMS logistics. There are two main approaches to solving RL problems. Model-free RL strategies, such as Q-Learning, proximal policy optimization (PPO), and actor-critic methods such as A3C, enable agents to learn optimal ambulance routing or stationing policies through interaction with dynamic environments. However, model-based approaches like MuZero and World Models simulate EMS scenarios, allowing preemptive decision-making under uncertain conditions. There is also a hybrid, actor-critic approach, which employs both value functions and policy search. These methods are particularly valuable in high-stakes, resource-constrained contexts where adaptability and timely action are paramount. Taken together, these AI paradigms form a multi-domain architecture capable of enhancing both the efficiency and quality of care in prehospital emergency services. More recently, reinforcement learning from human feedback (RLHF) has gained prominence, particularly in training large language models (LLMs) [18], where fine-tuning AI models is trained using expert demonstrations and human preferences to match automated decision-making with ethical norms and clinical priorities.

2.3. Literature Search Strategy

Our systematic review approach conformed to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) criteria [19], resulting in a complete and repeatable research procedure. A comprehensive literature search was carried out using several online databases to guarantee the complete retrieval of pertinent articles from 2018 to March 2025. The following databases were searched: IEEE Xplore, Web of Science, Scopus, and PubMed. Furthermore, a manual screening of reference lists from included studies and relevant reviews was conducted to find any further applicable research. The search method utilized a combination of keyword-based queries and Medical Subject Headings (MeSH) related to AI and ML in prehospital and EMS, ensuring a comprehensive and systematic exploration of the relevant literature. The complete list of search terms is provided in Table 1. This review's methodological strategy encompassed several components, including the formulation of a search strategy, the application of eligibility criteria, the study selection process, the identification and classification of AI domains, the algorithms used, prehospital areas of application, clinical subclasses, explainable AI techniques, and the data types, categories, and relevant features considered across the included studies. The process involved three independent reviewers. In order to minimize the possibility of bias in study inclusion and reporting, reviewers 1 and 2 carried out the PRISMA-based selection and screening separately and simultaneously. In order to ensure consistency and methodological rigor, a third reviewer was responsible for validating the overall research methodology.

Table 1. Search Terms Used in PubMed and Related Databases.

Concept	Keywords and Synonyms
Artificial Intelligence	"Artificial Intelligence" OR "AI" OR "Machine Learning" OR "Deep Learning" OR "Natural Language Processing" OR "Large Language Models" OR "GPT" OR "BERT" OR "LLM"
Prehospital Emergency Care	"Prehospital" OR "Emergency Medical Services" OR "EMS" OR "Out-of-Hospital" OR "Paramedic" OR "Ambulance" OR "Emergency Call"
Triage and Dispatch	"Triage" OR "Dispatch" OR "Emergency Dispatch" OR "Triage System" OR "Prioritization" OR "Severity Assessment"
Clinical Tasks	"Diagnosis" OR "Risk Assessment" OR "Clinical Decision Support" OR "Care Pathway" OR "Patient Stratification"
Data Modalities	"Audio Data" OR "Free Text" OR "Clinical Notes" OR "Dispatch Text" OR "ECG" OR "Time Series Data" OR "Multimodal Data"
Clinical Outcomes	"Mortality" OR "ICU Admission" OR "Hospital Admission" OR "Response Time" OR "Time to Treatment"

2.4. Study Selection

The initial comprehensive literature search across three main databases (PubMed, IEEE Xplore, and Web of Science) yielded a combined total of 1077 unique records (PubMed: 929; IEEE Xplore: 138; Web of Science: 113). Additionally, nine articles were identified using the ArticleRadar method. The screening process followed a structured three-stage approach—beginning with title screening, followed by abstract examination, and concluding with full-text review conducted in three rounds—resulting in the exclusion of 849 articles from PubMed, 127 from IEEE Xplore, and 97 from Web of Science, based on the criteria described in Figure 4. After deleting 60 duplicates, a full-text analysis was then performed on the remaining 98 studies after initial screening, excluding 34 of them because they did not meet all eligibility criteria for the following reasons (19 lacked focus on the specific application area of prehospital emergency care, 9 did not present empirical data, and 4 were removed because they were retrospective studies without prospective relevance). Consequently, 54 studies were included in our review. In addition, by examining the reference lists of included studies of LLM-based methods in EMS, we identified a further 11 studies, bringing the total number of studies included in this review to 65. The resulting studies reflected the following distribution across these databases: 43 from PubMed, 8 from IEEE Xplore, and 14 from Web of Science. Tables 2–4 provide a comprehensive overview of artificial intelligence (AI) applications in emergency medical services (EMS), systematically categorized into four primary domains, as follows: diagnostics, risk assessment and care pathways, triage and prioritization, and other applications. Each table summarizes the studies by indicating the country of origin, the types of data employed (structured, unstructured), the number of patients or cases analyzed, the computational approach used, and the relevant performance metrics. Collectively, these tables capture the diversity of AI methodologies—ranging from machine learning classifiers to advanced deep learning architectures—applied across varied prehospital contexts and healthcare systems. They also highlight differences in geographical adoption, dataset characteristics, and clinical focus areas.

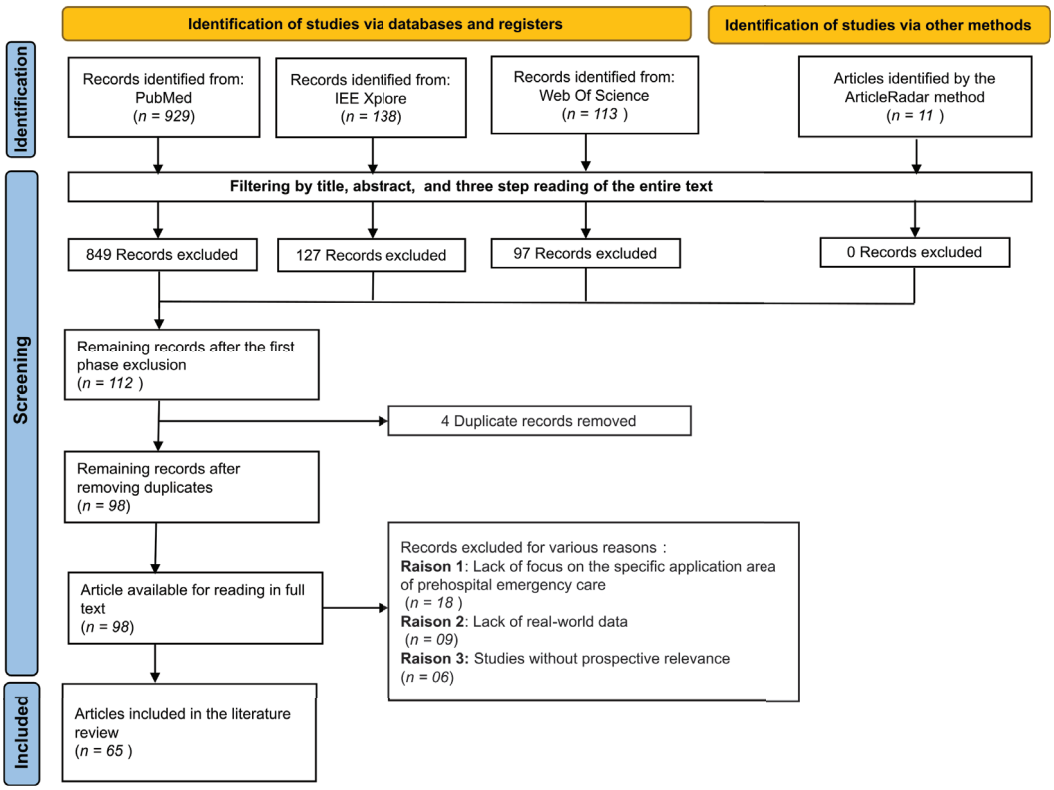


Figure 4. PRISMA flowchart of the results of the literature search.

Table 2. Summary of AI applications in EMS studies: Triage and Prioritization.

Ref.	Task/Outcome	Data	Nb Patients	Pays	Applied ML Algorithms	AUROC	XAI	SD	UD
[20]	Dept. ID Probability	Triage Records Data	200,000	China	DL (Bi-LSTM)	0.85	No	No	Yes
[21]	Ambulance dispatch prioritization	Dispatch Text	11,971	Australia	ML with NLP	0.94	No	No	Yes
[22]	Undertriage	Call Transcripts	19,114	Japan	ML (RF and XGB)	0.81	No	No	Yes
[23]	Emergency Reason Categories	Audio Data	1,982,746	Spain	DL/NLP (BERT)	0.95	No	Yes	Yes
[24]	Hourly No. of admissions	ED Data	87,600	France	ML(XGB, RF, LightGBM)	0.93	No	No	Yes
[1]	Emergency calls' reason	EMS data	888,469	France	NLP/DL(GPT-2)	0.80	No	No	Yes
[25]	Life-Threatening Level, Emergency System Jurisdiction	EMS Data	722,270	Spain	NLP/DL(DNN)	0.759	No	Yes	Yes
[26]	Emergency calls severity	Audio data	904	France	NLP/DL(BERT)	0.71	No	No	Yes
[27]	Emotional Content	Audio Data	337	Taiwan	ML(SVM)	0.94	Yes	Yes	Yes
[28]	Emergency Severity Index	Patient Outcome Data	534,730	Uganda	ML(XGBoost, RF)	0.7	Yes	Yes	No

Table 3. Summary of AI applications in EMS studies: Risk Assessment and Care Pathways.

Ref.	Task/Outcome	Data	Nb Patients	Pays	Applied ML Algorithms	AUROC	XAI	SD	UD
[29]	Preventable ED Admission	Hospital Admission Records	379,874	UK	ML (XGBoost, AdaBoost, RF)	0.83	No	Yes	No
[30]	Inappropriate Transfer	Patient Records	94,256	South Korea	ML(SVM, XGBoost, DT)	0.813	No	No	Yes
[4]	Coronary Angiography	Care Workflow Data	2398	USA	ML/DL(FCN, LR, XGBoost, SVM)	0.88	No	Yes	No
[31]	Travel Time	Ambulance Travel Data	406	Mexico	ML(RF)	0.69	No	Yes	No
[32]	Stroke Admission Delay	Stroke/Transport Data	450	China	ML(SVM)	0.846	No	Yes	No
[33]	Daily Ambulance Demand	Ambulance Demand Data	-	Singapore	ML/DL(DNN,LightGBM)		No	Yes	No
[11]	Critical Care	EMS Critical Care Data	8,981,181	South Korea	DL(DNN)	0.867	No	Yes	No
[34]	Hospital Admission	Risk Score Data	38,203	Sweden	ML XGBoost	0.89	No	Yes	No
[35]	Association Probability	ED Data	14,032	USA	ML(LR)	0.87	No	Yes	No
[36]	Patient admissions to ED	EMS Data	722,270	France	NLP/DL (Transformer)	0.759	Yes	Yes	Yes
[37]	Ambulance Demand	Ambulance Demand Data	-	Taiwan	DL/ML(ANN, SVM)		No	Yes	No
[38]	Cluster Detection	Patient Outcome Data	1510	South Korea	ML(K_means,BDSCAN)	0.95	No	Yes	No
[39]	1-Month Mortality	Cardiac Arrest Outcome Data	4335	Japan	ML(LR,SVM, RF,DNN)	0.89	No	Yes	No
[40]	1-Month Survival	Survival Data	45,000	Sweden	ML(RF)	0.86	No	Yes	No
[41]	New Injury Severity	Trauma Registry	75,602	Sweden	ML(XGBoost)	0.88	No	Yes	Yes
[42]	Stroke Encoding	Stroke Data	15,365	China	DL/NLP (LSTM, Transformer)	0.83			
[43]	ED Discharge Prediction	Hospital Admission Data	584,925	Canada	ML	0.77	No	Yes	No
[44]	Life-Threatening Event Classification	EMS Data	1,982,746	Spain	DL/NLP	0.71	No	No	Yes
[45]	Subsequent Event	EMS Data	11,846	Finland	NLP/DL(FastText)	0.66	No	No	Yes

Table 4. Summary of AI applications in EMS studies: Diagnostics and Other Applications.

Ref.	Task/Outcome	Data	Nb Patients	Pays	Applied ML Algorithms	AUROC	XAI	SD	UD
[46]	Traumatic Cardiac Arrest	Patient Transport Data	1,156,584	USA	ML/DL(DNN, SVM)	0.95	No	Yes	No
[47]	Anaphylaxis Event	Medical safety data	516	USA	NLP/ML (LSTM, Transformer, DT)	0.7	No	No	Yes
[48]	Stroke Mimic	Triage Records Data	402	China	MLP with LASSO selection	0.87	No	Yes	No
[49]	Cerebral Prognosis	Resuscitation Data	8274	Japan	DL(DNN)	0.94	No	Yes	No
[50]	Acute Coronary Syndrome	Clinical Data	555		ML (SVM, MLP)	0.84	No	Yes	No
[51]	Acute Aortic Syndrome	Triage Records Data	976	France	ML(LR)	0.87	No	Yes	No
[45]	Non-Transport Events	EMS Decision Data	11,846	Finland	ML(RF, SVM)	0.66	Yes	Yes	No
[52]	Major Trauma	Dispatch Data	114	Taiwan	NLP/ML (TF-IDF, NBayes)	0.65	No	No	Yes
[53]	Intracranial Hemorrhage	Triage Records Data	2123	Japan	ML(XGBoost, RF, SVM)	0.8	No	Yes	No
[12]	Out-of-Hospital Arrest	Audio Data	820	France	NLP/ML(LR, RF, DNN)	0.75	No	No	Yes
[54]	Stroke	Clinical Data	1446	Japan	ML(RF, SVM, XGBoost)	0.89	No	Yes	No
[27]	Emotional Content	Audio Data	337	Taiwan	ML(SVM)	0.94	No	Yes	No
[55]	Large Vessel Occlusion	Stroke Data	6305	Japan	ML(RL, SVM, RF)	0.89	No	Yes	No
[56]	Out-of-Hospital Arrest	Emergency Call Data	108,607	Denmark	DL(ARS)	0.98	No	Yes	No
[57]	Out-of-hospital cardiac arrest	Dispatch Data	3573	Canada	ML(LR, DNN)	0.87	No	Yes	No
[58]	Classification of emergency record categories (ICD-10)	Chinese EMS records	3500	China	DL/NLP(DNN,BERT)	0.82	No	Yes	Yes
[59]	Mortality Prediction	Trauma Data	191,191,249	USA	ML (LR,XGBoost)	0.85	No	Yes	No
[60]	Stroke Prediction	Helpline Call Data	2,361,178	Denmark	DL/NLP	0.98	No	No	Yes
[46]	In-Transit Arrest	Transport and Patient Data	460,521	Japan	ML/DL (SVM, ANN)	0.95	No	Yes	No
[61]	Infarction Prediction	Nationwide Cardiac Data	184,557	South Korea	ML(XGBoost, GB)	0.82	No	Yes	No
[62]	Consciousness Level	Triage Records Data	460,865	USA	ML/DL(LR,RF,DNN)	0.89	No	Yes	No
[63]	Sepsis, heart attack or cardiac arrest	Audio Data	71	South Africa	ML(DT, XGBoost)	0.82	No	Yes	Yes

Table 4. *Cont.*

Ref.	Task/Outcome	Data	Nb Patients	Pays	Applied ML Algorithms	AUROC	XAI	SD	UD
[64]	Treatment Adherence	Trauma Records	142	USA	NLP methods	0.70	No	Yes	Yes
[65]	Readmission Prediction	Cardiac Arrest Outcome Data	256,878	USA	ML/NLP	0.9	No	No	Yes
[66]	Detection of ST elevation myocardial infarction	ECG Data	371	Sweden	DL(ANN)	0.90	No	Yes	No
[67]	Electro-Mechanical Rhythm	ECG Data	1140	USA	ML(SVM)	0.93	No	Yes	No
[62]	Survival Prediction	Triage Records Data	460,865	South Korea	ML(RF), DL(ANN)	0.89	No	Yes	No

3. Benefits of AI in EMS

3.1. Predictive Modeling for Critical Care

The application of artificial intelligence (AI) in emergency call processing and triage has transformed prehospital decision-making by increasing response efficiency and accuracy. For example, ref. [20] developed an intelligent triage model using a Bi-LSTM network with character embeddings with the goal of optimizing telehealth services by effectively classifying patient complaints. Similarly, addressing the issue of undertriage, ref. [22] constructed machine learning models to increase telephone triage accuracy and mitigate undertriage hazards. They trained five models using data from Japan's after-hour house-call service, with random forest and XGB having the highest AUROC scores (0.81 and 0.80, respectively). The study demonstrates how incorporating machine learning into triage methods improves patient safety and results. Ref. [53] developed a prehospital triage system to identify severe cerebral hemorrhage utilizing numerous machine learning techniques. The extreme gradient boosting (XGBoost) model has the best predictive performance, with an ROC-AUC of 0.80 and a PR-AUC of 0.46. This technology proved effective at identification using just pre-transportation data, allowing ambulance personnel to make better judgments about patient route and treatment. Similarly, ref. [55] created the Japan Urgent Stroke Triage Score, which is an improved prehospital stroke categorization method. By using logistic regression, random forests, and XGBoost on multi-center cohort data, they achieved AUCs of 0.89 for large vascular occlusion, with performance higher than that of existing prediction models. These findings emphasize the potential of ML to improve the early diagnosis of life-threatening illnesses, allowing for faster interventions. Furthermore, ref. [48] developed a multilayer perceptron (MLP)-based model to improve prehospital triage by detecting stroke mimics (SMs) in patients suspected of having a stroke. In a retrospective examination of 402 individuals, 20.5% were recognized as SMs while hospitalized. The model performed admirably, with an area under the ROC curve (AUC) of 0.855 in the external testing set, exceeding established grading methods. These findings show the potential of AI-driven triage systems to improve prehospital decision-making and expedite stroke patients' treatment pathways.

Recent advances in artificial intelligence (AI) have considerably improved prehospital diagnosis, notably via the creation of machine learning models to assist emergency medical services (EMSs) in making fast and precise decisions. For instance, ref. [46] created a machine learning model that used a random forest approach to predict cardiac arrest during the transport of trauma patients. This model has an area under the curve (AUC) of 0.95 and a negative predictive value of 0.99, indicating its potential for detecting patients at risk of cardiac arrest during transportation. Similarly, ref. [47] used natural language processing and machine learning approaches in electronic medical records to enhance the detection of anaphylaxis patients. Their model has a cross-validated AUC of 0.70 in development data and 0.67 in external validation data, demonstrating improved detection skills for anaphylaxis incidents. Ref. [54] created a machine learning-based diagnostic method that identified strokes with an AUC of 0.980. This excellent predictive value

demonstrates the algorithm's potential for early and accurate stroke detection in prehospital situations. Furthermore, ref. [27] constructed a machine learning model that uses text from emergency medical dispatch messages to estimate the degree of patient injury. The model performed similarly to experienced dispatchers, indicating its usefulness in early detection of severe incidents to enhance resource allocation and patient outcomes. Ref. [49] created a deep neural network to predict neurological outcomes in individuals with OHCA. After analyzing data from 8,274 instances, the model obtained an AUROC of 0.94, suggesting a great predicted accuracy. The study found non-linear correlations between transport time and defibrillation or medication administration time, implying that improving these variables could improve emergency medical services' performance in OHCA scenarios. Meanwhile, the application of Named Entity Recognition (NER) in prehospital trauma treatment has been studied as a means to improve the extraction of crucial information from clinical narratives. By correctly detecting and categorizing critical medical phrases, NER systems can improve the efficiency and accuracy of data processing in emergency medical services [68].

3.2. Optimized Resource Allocation

Data analytics and AI significantly enhance resource allocation in EMS by optimizing response times through advanced analysis of spatial and temporal data. By integrating GPS data with historical traffic patterns, AI-driven models can predict potential delays. Ref. [31] demonstrated that by predicting the best routes and identifying any delays, EMS teams can effectively handle crowded areas. Ref. [33] developed a machine learning model based on ambulance demand data to predict ambulance demand, facilitating preallocation and reducing response times. These data-driven approaches, which make use of both historical and real-time information, have a strong connection with better patient outcomes, especially in time-sensitive emergencies where immediate action is essential, such as cardiac arrest or stroke. Ref. [55] demonstrated that using reinforcement learning further improves these systems by dynamically adjusting resource allocation and routing strategies according to real-time situations, guaranteeing EMS efficiency throughout a variety of operational conditions. Similarly, Ref. [34] used dispatch and ambulance data to create machine learning (ML)-based models that predicted the probability of hospital admission, critical care demands, and two-day mortality in prehospital emergency medical occurrences. Their models outperformed standard rule-based algorithms and human prioritizing judgments, with higher concordance indices and stronger performance in prospective validation datasets.

3.3. AI in Triage and Diagnostics

Emergency physicians have emphasized the important need for AI technologies [69], notably in documentation, clinical prediction, and vital monitoring, claiming that these breakthroughs may dramatically enhance workflows and patient care. They have also stressed the need to improve handoffs among emergency physicians [70]. This emphasizes the need to include physician involvement in the creation of AI technologies to enable their appropriate incorporation into clinical practice. Similarly, ref. [25] developed DeepEMC2, a deep ensemble multitask model that uses multimodal data to classify emergency medical call situations based on their life-threatening nature, permissible response latency, and emergency system jurisdiction. This model outperformed the Valencian emergency medical dispatch service's in-house triage routine, with significant increases in macro F1 scores across all classification tasks. Ref. [44] applied deep continuous learning to identify life-threatening events from emergency medical call episodes based on free-text dispatcher observations. This method enabled the model to adapt to new data over time, ensuring

stability and great performance in detecting life-threatening scenarios. These studies highlight the abilities of advanced machine learning and deep learning approaches to improve decision support systems in emergency medical services by improving accuracy, response times, and overall patient outcomes.

3.4. LLMs in EMS and Call Regulation

The use of large language models (LLMs) in EMS is likely to transform emergency call regulation by improving decision support and real-time data processing (see Table 5). Recent advances, such as the Chu et al. [71] professional agents (PAgents) framework and the Tang et al. (2023) [72] Multi-disciplinary Collaboration (MC) framework, show the disruptive potential of LLMs in carrying out specialized, professional-level activities within the healthcare sector. These new frameworks demonstrate the ability of LLMs to perform zero-shot reasoning and facilitate collaborative multi-agent discourse, both of which are critical in high-stakes contexts like EMS, where rapid and accurate decision-making is required. Nonetheless, considerable obstacles exist in the area of call regulation. Ref. [73] reported that ChatGPT and Copilot outperformed traditional nurse-led triage in detecting high-acuity patients in prehospital settings. They highlighted the need to integrate real-time emergency department (ED) capacity data to minimize overcrowding and ensure the effectiveness and reliability of emergency services. The two major barriers are the limited volume and specificity of medical training data—constrained by cost and privacy concerns—and the requirement for deep domain expertise combined with strong reasoning capabilities to address complex medical questions. To address these issues, Li et al. (2023) [74] described the specialized model ChatDoctor, which illustrates the usefulness of fine-tuning LLMs with domain-specific data and implementing real-time information retrieval techniques, hence improving the accuracy and reliability of medical advice. Guo et al. (2024) [75] noted that continual improvement and refinement of these models are critical for overcoming present hurdles and realizing the full potential of LLMs in EMS call control.

Table 5. Summary of EMS-related LLM studies using structured data (SD), unstructured data (UD), and explainable AI (XAI).

Ref.	Year	Country	Data	Clinical Area	Task	Algorithm	Performance	XAI	SD	UD
[76]	2024	Israel	Triage Records Data	Triage and Prioritization	Emergency Severity Index	GPT-4	GPT-4 vs. human	No	Yes	No
[77]	2024	Italy	Free-text ED records	Diagnostic	Injury type categories	GPT-4	No	No	No	Text
[78]	2023	USA	Chatbot Data	Care Pathways	Documentation	GPT-3.5	No	No	No	Text
[79]	2024	USA	Free-text medical communication scenarios	Triage	Emergency Detection	LLaMA	0.98	No	No	Text
[80]	2025	USA	Ambulance Demand Data	Triage and Prioritization	Triage and Dispatch Prioritization	GPT4o-Mimi	0.76	No	No	Text
[81]	2024	USA	MIMIC-III-ED	Diagnostic	Diagnostic Errors	ClinicalBERT	0.77	No	Yes	Text
[82]	2024	USA	EHRs Data	Diagnostic	GPT Diagnostic Evaluation	ChatGPT-3.5/GPT-4	0.81	No	No	Text
[73]	2025	USA	EMS Data	Triage and Prioritization	EMS Triage and dispatch Evaluation	ChatGPT-3.5/GPT-4	0.87	No	Yes	Text
[83]	2024	Turkey	Triage Records Data	Triage and Prioritization	Emergency Severity Index Prediction	ChatGPT-3.5/GPT-4	0.94	No	Yes	Text
[84]	2024	USA	ED Data	Triage and Prioritization	Acuity levels Classification	ChatGPT-3.5/GPT-4	0.89	No	No	Text

Recent studies underscore the diverse potential of LLMs. Ref. [76] compared the performance of the GPT-3.5 and GPT-4 models to human physicians to generate medical

scores. This study found that while LLM models showed a reasonable level of agreement with human experts, they were unable to replicate complex clinical judgments. These results highlight the necessity of including physician input in AI architecture to facilitate successful integration into clinical practice, especially in high-stakes scenarios such as emergency departments.

Ethical Considerations and Integration into EMS Workflows

A strong ethical foundation is necessary for the appropriate application of LLM in emergency medicine. Compliance with established privacy and data protection standards, such as HIPAA or the General Data Protection Regulation (GDPR)—to protect patient data—is an essential consideration, especially when integrating LLM-based agents into emergency call analytics and other ED applications [85]. Implementing strong data protection measures is critical to safeguarding sensitive patient information, as these models are integrated into workflows that process sensitive patient information. To secure patient trust by adhering to strict confidentiality regulations and preventing potential breaches. Furthermore, fairness and bias in LLMs in healthcare also raise crucial ethical issues [86]. Moreover, explainability and transparency are equally important, serving as fundamental principles that direct the integration of LLMs into EMS systems to support decision-making by dispatchers and physicians. Hallucinations—situations in which an LLM produces inaccurate or fake content—pose a serious concern in emergency medical situations when decisions could have an immediate impact on the lives and safety of patients. Studies suggest numerous strategies to reduce hallucinations and improve factual accuracy, strategies like retrieval-augmented generation (RAG) [87], agentic AI [88], and prompt engineering [89]. By combining these approaches, LLM-based systems can offer more reliable evidence-based assistance to EMS professionals and patients.

4. Discussion

The tremendous growth in scientific publications related to AI across the prehospital sector reflects a broader trend in medical research where data-driven decision-making is becoming increasingly critical.

The absence of a universal AI model in prehospital applications, as revealed by our analysis, underscores the importance of context-specific modeling approaches. This finding is consistent with previous studies on AI in emergency medical services that highlight the uniqueness of each AI model based on its data and predictions. Our review expands on this comprehension by demonstrating how AI model selection for the prehospital care area must be customized to specific data topologies and healthcare contexts, frequently requiring direct input from healthcare workers through field interviews to ensure model accuracy, as well as identifying the evaluation metrics that are the most suitable for AI models in prehospital emergency medical services.

4.1. Publication Countries

Figure 5 displays the papers included in the analysis, organized by the country in which they were developed. A large number of research articles were published in the United States (23.1%), followed by Japan (12.3%) and France (9.2%). China and South Korea each represented 7.7% of the studies, while Taiwan contributed 6.2%. Sweden and Spain accounted for 4.6% each, and Australia, Finland, Canada, and Denmark each contributed 3.1%. The remaining countries—including the United Kingdom, Mexico, Singapore, Uganda, South Africa, Israel, Italy, and Turkey—each represented 1.5% of the total. The total value of the table exceeds the number of articles studied because some publications originated from multiple countries. Due to the growing need for rapid

response systems and the frequent occurrence of pandemics and natural disasters, countries such as the US and Japan have demonstrated strong interest in using AI in prehospital emergency medical services (EMS). Several nations, particularly those with well-established healthcare systems—such as the United States, the United Kingdom, and France—produced a greater number of publications. These countries possess greater financial resources and research capacity, enabling them to dominate scientific production in this area of interest.

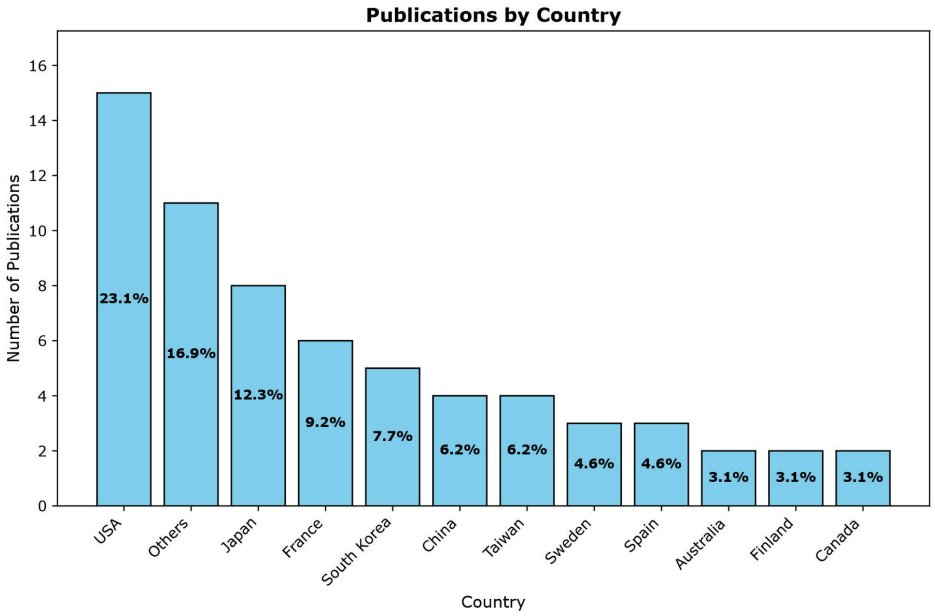


Figure 5. Distribution of scientific publications by country.

4.2. Stage-Wise Prehospital Distribution

The integration of AI in prehospital emergency medical services (EMS) covers multiple stages, ranging from the initial call to arrival, as illustrated in Figure 6. Regulation is the most common focus, appearing in 73.33% of studies, indicating a significant emphasis on compliance and operational procedures. Clinical assessment comes in second, at 58.33%, highlighting the importance of clinical applications in triage and diagnosis, with 45% and 41.67% of studies focusing on orientation and decision-making, respectively, highlighting efforts to improve clinical decision-making and patient transportation. Call assistants and ambulatory care are less frequent, with 16.67% and 20%, respectively, showing possible research gaps in call handling and ambulatory care optimization.

The diversity of AI models (ML, DL, NLP) identified in our review, which range from supervised learning to natural language processing, shows the maturity of the field and the various needs of healthcare organizations. Their sustained coexistence demonstrates that various healthcare situations may benefit from a variety of AI approaches, combining generative models and classical methods. This highlights the possibility of collaboration across distributed and agentic AI systems, which may integrate multiple models and customize solutions to individual demands. This study has important implications for future implementation efforts, underlining the value of adaptable and context-sensitive techniques in EMS.

Figure 7a displays the historical distribution of AI-based research in prehospital care by its clinical application. There is a recognized increasing trend, with the number of publications peaking in 2022 (n = 17) and then gradually dropping in 2023 and 2024. Diagnostic applications have been the most prevalent throughout the study period, with more than 35 studies, indicating a sustained interest in supporting early medical decision-making. After 2020, the creation of care pathways increased, and risk assessment reached its

peak in 2022 (n = 5). Interestingly, from 2023, triage and prioritization actions became more prevalent, demonstrating an increasing focus on operational effectiveness and real-time identification of patients in emergency scenarios.

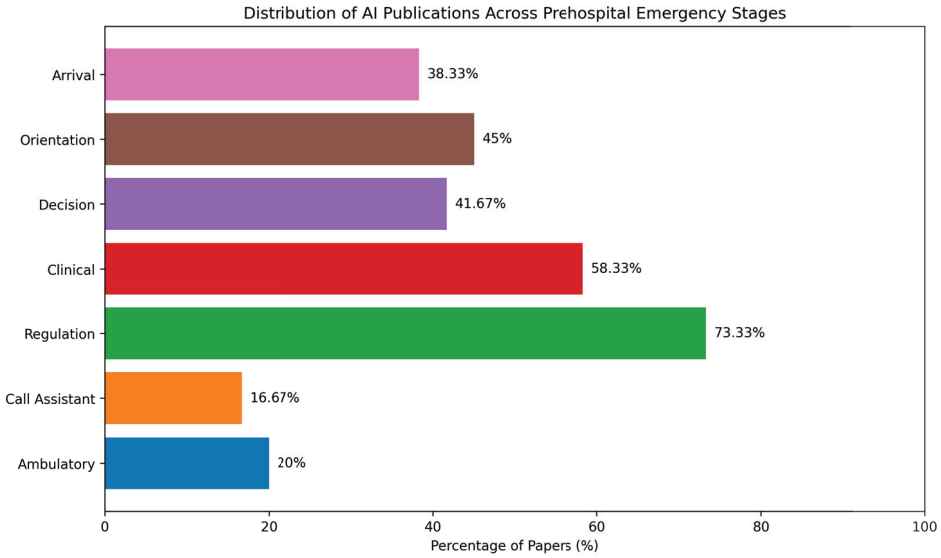
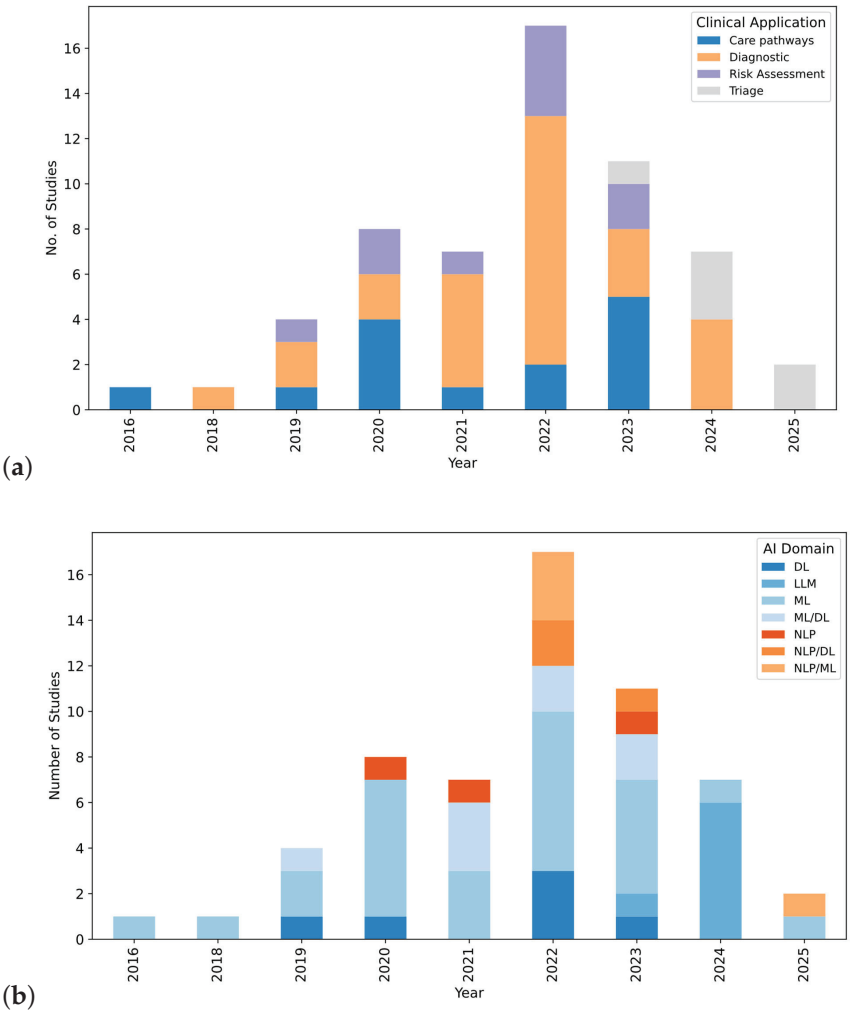


Figure 6. Distribution of AI Publications Across Prehospital Emergency Stages.



(**a**)

(**b**)

Figure 7. Distribution of studies by publication year; (**a**) prehospital clinical applications, (**b**) AI domain methods.

Figure 7b illustrates the types of artificial intelligence interventions in emergency medicine across these studies. Furthermore, we found that over 60% of the research relied on traditional machine learning (ML) and deep learning (DL) techniques, which accounted for the majority of the methodologies used, and 17 (20%) studies evaluated hybrid ML-NLP- and DL-NLP-based interventions. However, 2022 observed a shift in AI strategies when large language models (LLMs) like GPT-based architectures emerged. A total of 10 (15%) studies focused on LLMs. Notably, we observed LLMs were employed in four additional studies in 2023, continuing the trend. This development demonstrates a substantial methodological change toward language-aware models that can interpret free text input, which is especially pertinent to prehospital system communication, dispatch analysis, and triage activities.

Figure 8 shows the annual publication count of research, classified by the specific AI domains applied to prehospital tasks. Since 2018, machine learning (ML) approaches have constituted the predominant uses, with a significant surge observed between 2021 and 2022, presumably fueled by the heightened interest in dispatch and triage optimization during the COVID-19 pandemic. Natural language processing (NLP) and deep learning (DL) methodologies have gained importance, especially in research utilizing free-text data and sophisticated model architectures to improve performance. Some studies combined NLP with ML or DL techniques. More recently, since 2022, the emergence of large language models (LLMs), such as ChatGPT, has led to at least seven studies exploring their use in various prehospital AI tasks, often aiming to benchmark their performance against human expertise.

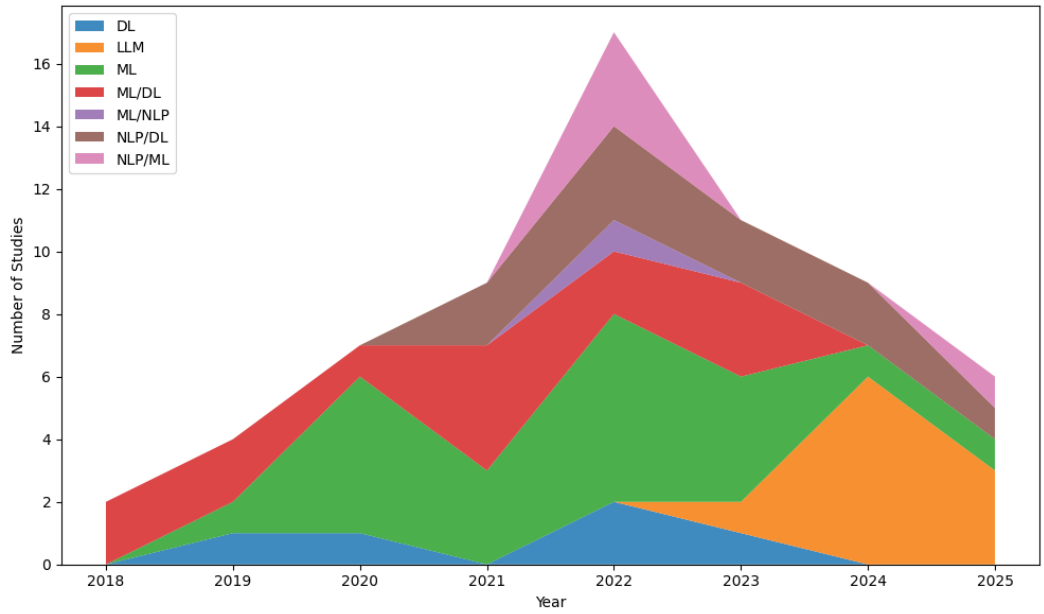


Figure 8. Stack plot of total publications per year, stratified by AI domain.

Figure 9 presents the frequency of different algorithms used across the reviewed studies in prehospital AI applications. Random Forest (RF) was the most commonly used algorithm among the 33 different algorithms that were found, appearing in 18 studies, or around 20.2% of all reported algorithm usages. Deep neural networks (DNNs) and support vector machines (SVMs) followed, each used in 15 studies (=16.9%). XGBoost (13 studies, 14.6%), logistic regression (LR), and decision trees (DT) were other frequently utilized techniques. Although they were less common, advanced techniques like Transformers, GPT, Llama, and LLMs are noteworthy because they have recently been used in clinical decision-making and natural language challenges. Although emerging LLMs like GPT-4,

LLaMA, and ClinicalBERT represent the changing environment of AI in prehospital care, traditional ML models continue to dominate.

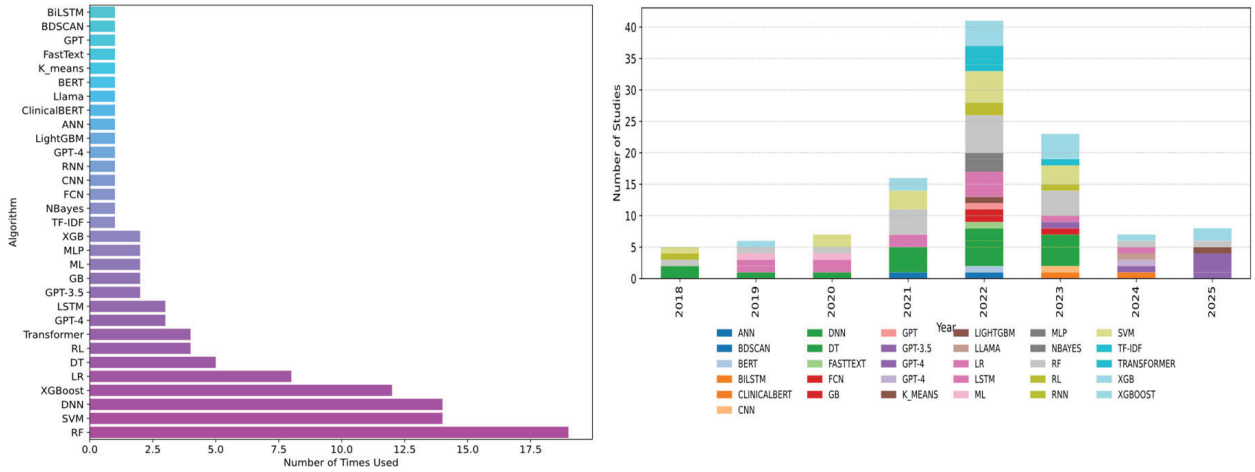


Figure 9. The frequency of different algorithms used across the reviewed studies in prehospital AI applications.

The collaboration between prehospital AI tasks, input data types, and the algorithms used is depicted in Figure 10. The prevalence of diagnostic applications, which use a variety of clinical, triage, and hospital datasets, emphasizes how important structured health records are to the advancement of AI. In order to meet the demands of real-time decision-making, triage, prioritization, and risk assessment activities increasingly rely on dynamic data sources such as dispatch logs, audio, and free-text communication. Across all areas, the most widely used machine learning models remain Random Forest, SVM, and DNN. A notable trend toward more context-aware, language-based intelligence is the integration of advanced NLP and LLM methods in recent research (e.g., GPT-4, ClinicalBERT), especially in text-rich tasks.

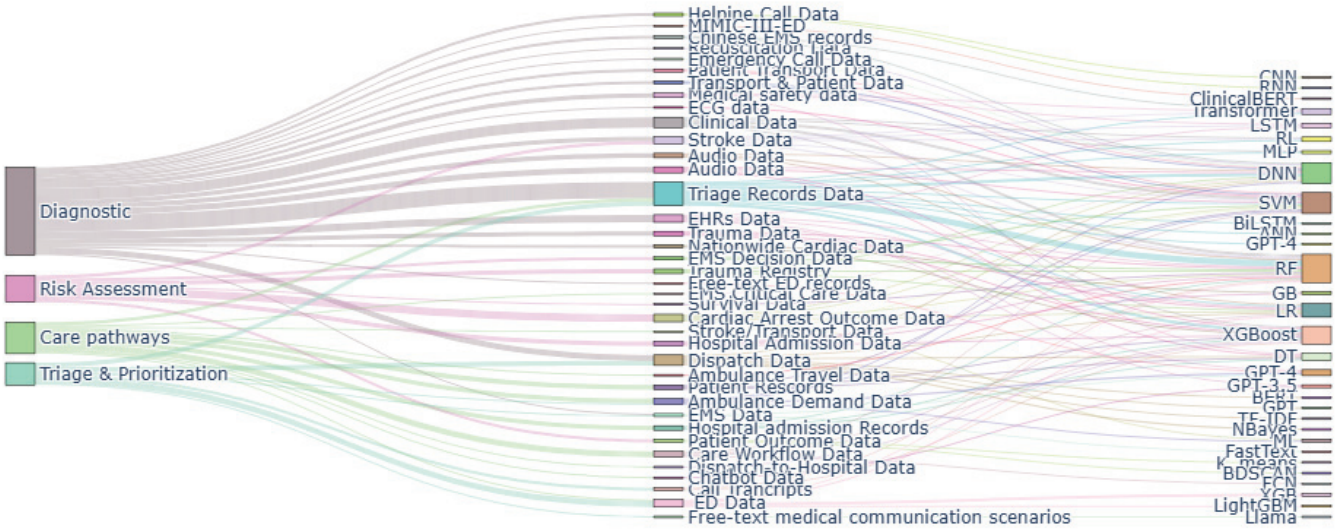


Figure 10. Summary of clinical purposes and data sources related to different ML models in AI-driven prehospital applications.

5. Database Integration and Data-Driven Approaches

Regarding the growing integration of data-driven technology into emergency medical services (EMS), there is still a lack of studies on large-scale, multimodal data management

in this field. Several crucial problems remain unresolved, and as digital health applications grow, new difficulties keep appearing. Key research issues are discussed in the parts that follow, such as managing both structured and unstructured modalities, scalable big data architectures, data privacy, data integrity, and data linkages across heterogeneous systems.

5.1. Big Data Architectures in EMS

The main aim of the solution proposed in this thesis is to improve the quality of emergency call regulation by providing better decision support and care. Recent experiments conducted as part of the SIA-REMU project in France have highlighted the need for architectures capable of continuously processing incoming calls, enriching the data with hospital sources (DMU, DPI), and structuring this information for top-down use, including generating dashboards for medico-operational management, producing quality indicators, and developing predictive triage models or emotional analysis via NLP [90]. We propose a system whose aim is to make all the data usable in decision support systems. For integrating different health data into one data warehouse, we need to first aggregate the EMS data together from different services. Typically, this architecture consists of four layers, namely, the data collection layer, data storage layer, data processing level, and finally, data analytics and the application layer, as shown in Figure 11.

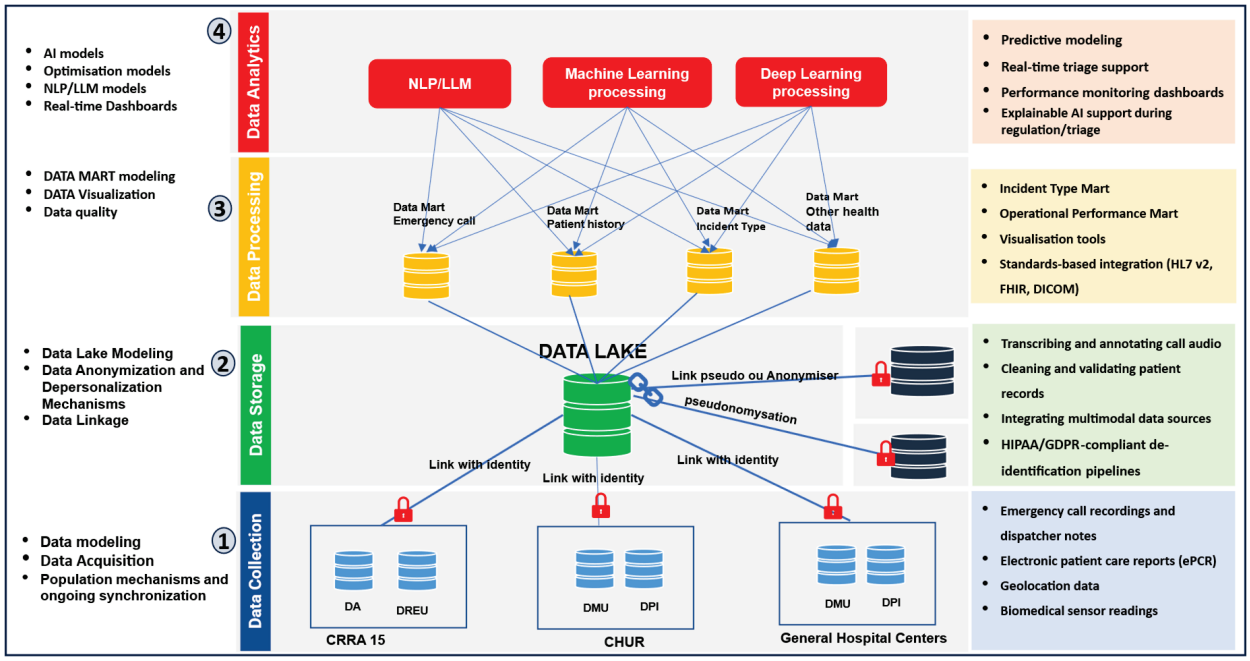


Figure 11. Four-level big data architecture for EMS.

5.1.1. Data Collection Level

The first level of the architecture—the data collection level—encompasses the secure centralization of data from regulation centers (calls, metadata, reports), hospital records (DPI, DMU), and regional systems (RPU, Dossier Communiquant de Cancérologie, etc.). Additionally, physiological data from patient sensors (blood pressure, pulse oximetry, and electrocardiogram) are becoming more widely accessible through connected devices, enhancing knowledge of the patient's condition immediately before hospitalization. This data is integrated using interoperability standards (HL7, FHIR) [91], while ensuring alignment with patient identities (or their pseudonymization in accordance with RGPD constraints).

5.1.2. Data Storage Level

This layer constitutes the foundational component of the architecture, where all collected data types are stored. Information obtained from the data collection layer can be categorized as structured, semi-structured, or unstructured. These datasets are subsequently ingested into the data lake, where they are later refined into data marts adapted to specific use cases, either for data analysis or AI applications for decision-making.

5.1.3. Data Processing Level

This is the central processing unit layer of the EMS data architecture, where collected raw data is refined, structured, and organized to support decision-making and predictive modeling. To meet management and analysis needs, data marts are created from the data lake. These marts organize the data into processing and decision-making categories, such as the mart of dispatched calls (response time, reasons, type of emergency), the mart of medical antecedents (diagnoses, emergency room visits), and the mart of on-site interventions (SMUR, transport). This architectural layer leverages distributed computing frameworks, such as Apache Hadoop or Apache Spark, to efficiently process and manage large-scale data volumes. In addition, this layer is responsible for orchestrating the procedures involved in transferring and storing relevant data into dedicated data marts for creating and implementing AI models, in order to facilitate tasks like emotion recognition, triage prediction, and call categorization.

5.1.4. Data Analytics and AI Application Level

This is the most advanced part of the architecture that produces intelligent services and applications for analysis, BI/reporting, optimization, and predictions. The goal of this layer is to extract relevant information from the processed data and convert it into knowledge that facilitates decision-making. This layer enables EMS stakeholders, such as medical regulators, emergency call assistants (ARMs), and healthcare administrators, to effectively interpret and assess the outcomes of analytics operations performed on emergency medical big data by means of interactive visual tools such as dashboards, maps, charts, and reports. Moreover, advanced AI techniques such as NLP, LLM, ML, and DL are used for continuous monitoring, efficiency evaluation, predictive modeling [33,58,64], and recommendations for resource optimization.

The main problem with emergency medical data is that 80% of it is unstructured data. A wide range of data is generated throughout the interactions with patients, encompassing clinical notes, diagnostic test results, treatment plans, and patient feedback, all of which contribute to a comprehensive understanding of patient care. Significant progress has been made in extracting relevant medical information from unstructured data for research purposes (see Table 6). These approaches for medical text data are generally manual, rule-based, and learning-based, including techniques such as natural language processing (NLP), computer vision (CV), and general machine learning and deep learning [61,65].

Structured data: Mostly utilized in 39 studies, structured data include triage records, trauma registries, hospital admission records [34,43], and patient transport data [31,33]. The Swedish Trauma Registry and South Korea's Nationwide Cardiac Data are two examples of databases that include tabular data with predefined fields, such as clinical outcomes and patient information [41,52].

Unstructured data: Used in 16 studies, unstructured data includes triage notes [20,36], free-text medical communication, and audio data (such as emergency call recordings from Denmark and France) [56]. These data formats are frequently processed using NLP and LLMs, such as BERT [58], GPT-4 [77], and Llama [79]. Furthermore, some studies [27,36]

combine structured and unstructured data to improve model performance using advanced deep learning methods or LLM approaches.

Table 6. Classification of structured and unstructured data in EMS with references.

Structured Data	References	Unstructured Data	References
Patient Demographics Information such as age, gender, and medical history that is stored in predefined formats and used for understanding patient profiles and predicting outcomes. Example: Forecasting EMS demand using patient demographics.	[25,29,45]	EMS Call Transcripts Audio recordings and transcriptions of emergency calls, which contain rich, unstructured information about the incidents. Example: NLP analysis of call transcripts to predict cardiac arrest.	[21,22,52,57]
Clinical Data Vital signs (e.g., heart rate, blood pressure), symptoms, and results from diagnostic tests like ECGs and blood tests, all of which are recorded in a structured manner. Example: Machine learning model for predicting prehospital delays using clinical data.	[46,48]	Clinical Notes Free-text notes written by EMS personnel and healthcare providers, describing patient conditions, treatments administered, and observations made during care. Example: Analyzing free-text notes using NLP to predict hospital admissions.	[45,92]
Operational Data Details about EMS responses, including dispatch times, response times, types of vehicles dispatched, and geographic locations of incidents, are all organized in structured databases. Example: Using operational data to optimize resource allocation and reduce response times by 16%.	[21]	Incident Reports Detailed narrative descriptions of EMS incidents, which often include unstructured text providing insights into the circumstances and actions taken during emergencies. Example: Analyzing incident reports to understand predictors of survival in out-of-hospital cardiac arrests.	[35,46,52]
Environmental Data Contextual information, such as weather conditions, traffic patterns, and special events, can influence EMS demand and response efficiency. Example: Integrating environmental data into machine learning models to predict ambulance demand.	[21]	Patient Feedback Unstructured feedback and comments from patients and their families provide qualitative data on the service experience and outcomes. Example: Analyzing patient feedback to improve EMS service quality and patient satisfaction.	[45]

The dataset illustrates several data categories employed in AI-driven emergency medical services (EMS) research, highlighting the complexity of prehospital care applications. Triage record data appears in 12 studies, applied to tasks such as stroke mimic prediction [76,83] and emergency severity index classification [48]; EMS and dispatch data, used in 8 studies, include ambulance demand data [31] and call transcripts for dispatch prioritization [21,57], as well as emergency detection [93]. Audio data, utilized in three studies, supports tasks like cardiac arrest detection based on phonetic characteristics [12] or emotional content analysis [27]. ECG and EHR data are used in particular situations, including pulse detection during cardiac arrest [39], classification of EMS records [58], and assessment of the correctness of GPT-generated diagnoses [82].

5.2. Data Privacy Issues

The management of data and fieldwork is intricate and necessitates striking a balance between research objectives, privacy, and secrecy, as well as a sensitive, transparent approach. According to HIPAA, there are eighteen different categories of personally identifiable information (PII) in medical data [94]. Processing health data is significantly hampered by the existence of PII in the data [95]. Consequently, a number of anonymization techniques have been put out to address this problem and make use of the patient data. In France, data security and privacy are essential considerations when using AI for personal health information, especially given the severe requirements of the General Data Protection Regulation (GDPR) and the Commission Nationale de l'Informatique et des Libertés (CNIL) [96]. The heterogeneity of data usage agreements, particularly regarding

the use of patient data to "train" AI models, poses substantial issues. Ensuring GDPR compliance while protecting patient data remains a critical challenge in the development and use of AI technology in healthcare.

5.3. Data Linkage

Effective monitoring and assessment of healthcare systems requires data linkage, which connects disparate datasets including prehospital records, emergency medical services (EMS) data, emergency department (ED) reports, hospital records [97], EMS electronic patient care reports (ePCRs), and emergency call data. Linked datasets offer a single view that improves management of patients and outcomes by enabling a comprehensive documenting of the patient pathway from the prehospital scene to in-hospital treatment. While evaluating many study issues, it is crucial to capture different stages of patient care (such as in- and out-of-hospital stays) [98]. Nonetheless, it is now challenging to match patient care at different stages due to stricter privacy laws governing medical information. Once a patient arrives at the hospital, data linking can give doctors instant access to prehospital data, which emphasizes the significance of data linkage. In an emergency, when fast and accurate data might impact crucial choices, this smooth information flow is especially important. The effective application of a machine learning algorithm to link EMS ePCRs with hospital EMRs is demonstrated in this paper [35]. The method dramatically improved the accuracy and timeliness of data availability, which is crucial for patient treatment, achieving high sensitivity (99.4%) and a positive predictive value (99.9%). Similarly, Miller et al. [92] demonstrated the significance of data linking by combining many datasets (ambulance, emergency department, hospital, etc.) to monitor trauma patients from the prehospital incident until discharge. This strategy improves patient care by increasing the availability and quality of prehospital data in hospitals, as well as allowing for improved monitoring and assessment of trauma systems.

5.4. Data Integrity

Data integrity is a cornerstone of effective emergency medical services (EMS), particularly when handling unstructured prehospital data such as dispatch call recordings, free-text triage notes, or audio-based symptom descriptions, which are often processed through techniques like character embedding or natural language processing for machine learning applications [30,56]. To ensure data accuracy, only authorized personnel (e.g., paramedics and dispatchers) should be able to make changes to records. This will prevent unauthorized individuals from making modifications or making mistakes that might affect clinical evaluations [47]. Inaccuracies, such as those caused by transcribing mistakes, background noise that muffles call audio, or incomplete entries made under pressure during field operations, are especially likely to occur in unstructured prehospital data [99]. These deficiencies can substantially impair the performance and reliability of predictive models for critical conditions, such as sudden cardiac arrest or severe trauma [21]. Moreover, there are new security and dependability issues brought about by the use of cloud-based EMS systems, which allow for real-time processing, storage, and remote access. Strict security measures, including cryptographic hashing, secure access control, and automatic error-detection methods, are necessary for such systems to guarantee that data is correct and verifiable from the time of collection until hospital transfer [77]. Ultimately, EMS still has a great deal of challenges in preserving data integrity since the industry must strike a balance between the operational necessity of prehospital care and the security measures required to guarantee that vital data are accurate, safe, and reliable for life-saving decision-making. This requires ongoing collaboration between technology experts, healthcare providers, and

policymakers to tackle all the data-related concerns that have been highlighted, including linkages, privacy, scalability, and multimodal data management.

5.5. Evaluation Metrics for AI Models in Prehospital Care

AI algorithms in prehospital care can possibly be developed as classification and regression problems, with evaluation metrics chosen based on their particular application. For classification-based AI models, which are often used in triage, risk assessment, and emergency call categorization, accuracy is a widely used metric that evaluates the proportion of correctly predicted cases out of the total cases. Moreover, accuracy, recall, and F1 score are significant classification metrics. In critical care decisions, for example, where an incorrect high-priority classification might lead to wasteful resource allocation, precision is crucial in preventing false positives. However, in other cases, such as identifying high-risk patients from emergency call transcripts, recall is crucial because failing to recognize a real emergency could have catastrophic repercussions. The F1 score, which is the harmonic mean of accuracy and recall, is especially beneficial in situations when both metrics must be balanced. Furthermore, the area under the receiver operating characteristic curve (ROC-AUC) and the equal error rate (EER) are utilized to assess model resilience in discriminating emergency severity levels.

Regression-based AI models, such as those used for forecasting response times, patient deterioration, or ambulance arrival times, frequently employ root mean square error (RMSE), mean absolute error (MAE), and R-squared (R^2).

Out of the 65 studies reviewed, 11 reported area under the curve (AUC) values above 0.90 [57,67,79], while 39 studies reported AUC values exceeding 0.80 for various outcomes such as life-threatening event classification [44], undertriage [22], out-of-hospital cardiac arrest [57], caller's emotional state [27], and stroke detection [54].

Traditional machine learning models such as RF and XGBoost demonstrated strong performance in some tasks like ICU admission [100] and cardiac arrest prediction during transport [46] (AUC 0.90–0.94). Advanced deep learning methods, such as Transformers, were often used with NLP techniques, which perform well on unstructured data, including injury type classification, survival prediction [62], and out-of-hospital cardiac arrest [42,57]. All these methods demonstrated strong performance, with AUCs often ranging from 0.75 to >0.90. Additionally, LLM approaches have demonstrated strong AUCs in tasks including Llama for early medical emergency identification (AUC = 0.98) [79]. GPT-based triage and dispatch prioritization [73]. Notably, comparisons between GPT-4's decisions and those of senior nurses and physicians underscore the potential of AI systems to match or exceed expert committees [76,80].

Figure 12 illustrates a collection of bubble charts that illustrate the performance (AUROC) of artificial intelligence models applied in prehospital emergency settings in several clinical subclasses. The performance of machine learning models in various clinical subclasses, such as care pathways, risk assessment, and triage and prioritization, is represented by each subplot, with distinct bubbles signifying published research. The visualization highlights attention to several interesting patterns. First, there are high-performing models in a number of countries (AUROC > 0.85), indicating that successful model development is not limited to a single geographic area. Interestingly, research from Sweden and China frequently shows high AUROC ratings, especially in the diagnostic and risk assessment categories. Furthermore, the distribution of studies by year indicates that 2023 and 2024 saw a rise in research effort, particularly in the Care Pathways category. This might be a sign of increased interest in using AI to improve patient routing and reduce unnecessary hospital admissions. Recent studies have focused on evaluating how well large language models (LLMs) such as GPT-4, GPT-3.5, ClinicalBERT, and Llama perform on a range of

emergency medical activities, utilizing either unstructured free-text input or a combination of free-text data and structured feature vectors. The evaluation strategies used include human-level comparisons, such as evaluating GPT-4 performance against human triage decisions, as well as traditional machine learning evaluation strategies Table A1. Strong model potential is indicated by high AUROC scores for tasks such as diagnostic generation (0.82 with GPT-4) and emergency detection (0.98 with Llama). However, some applications, such as categorization or documentation, exhibit little validation. All things considered, LLMs show promise in triage and dispatch situations; however, task type, dataset quality, and assessment design affect performance.

There is a very unequal distribution of research on the uses of large language models (LLMs) in EMS, as shown in Table 5. Most of the research that has been found comes from the US, with a smaller amount from other nations; Europe has not made much of an impact either. This imbalance reflects not only differences in research investment and adoption of innovation, but also systemic and regulatory barriers that uniquely constrain the integration of LLM-based on EMS in Europe. This imbalance reflects not only differences in research investment and innovation adoption but also several systemic factors, such as limitations in data availability, persistent interoperability issues, and governance challenges, as discussed above.

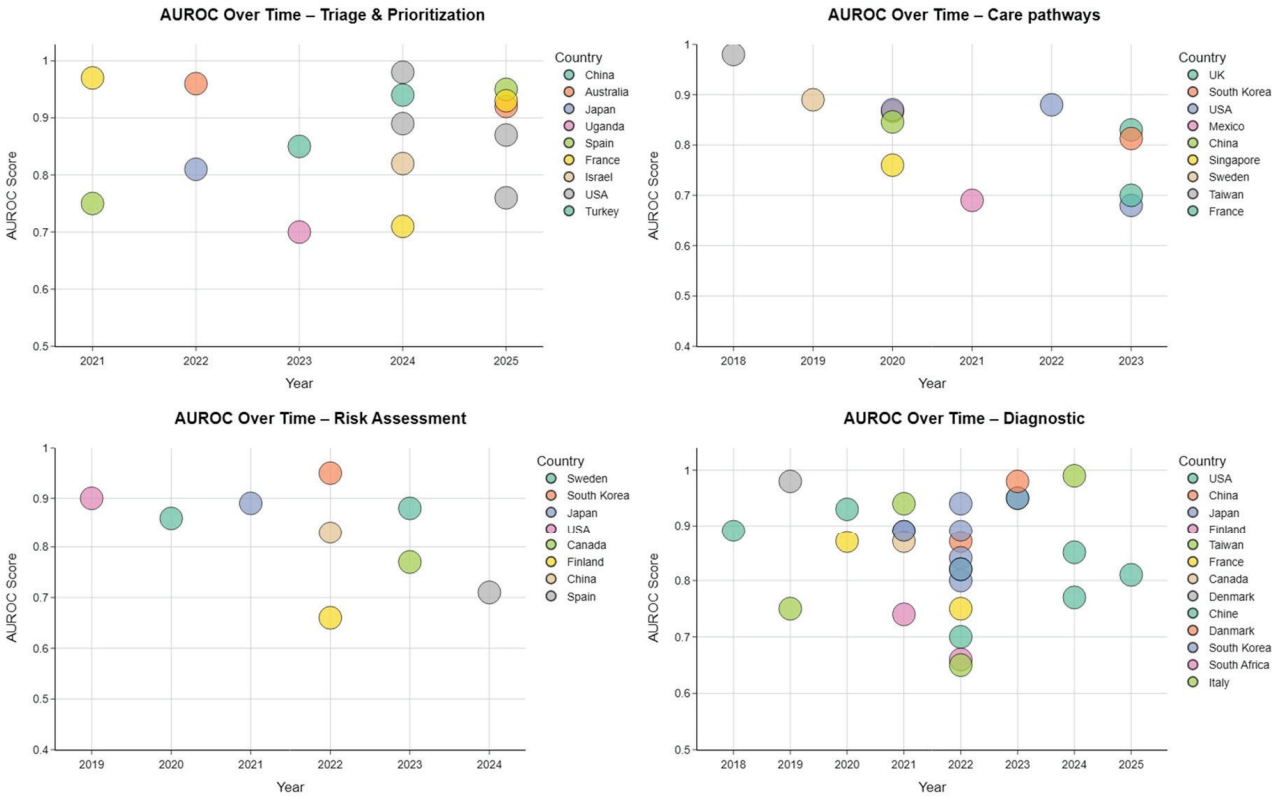


Figure 12. The performance (AU-ROC) of artificial intelligence models applied in prehospital emergency settings across several clinical subclasses (diagnostics, triage and prioritization, risk assessment, and care pathways).

5.6. Lack of Explainable AI in EMS Models

5.6.1. Explainable AI in the EMS Context

Explainable artificial intelligence (XAI) has become a crucial factor in emergency medical services (EMS) for the implementation of safe and reliable AI. XAI techniques in EMS should ideally give a detailed description of a system's strengths and weaknesses, explain the logic behind previous results, explain the analytical procedures being used now,

and provide the justification for upcoming suggestions. These kinds of skills are especially crucial for prehospital decision-making, as physicians and dispatchers have to make quick judgments under pressure. By revealing the underlying reasoning behind predictions, such as resource allocation, ambulance dispatch optimization, or priority classification of events, XAI can improve adherence to accepted clinical standards and make it easier to audit AI-assisted decisions in real time [101], facilitating oversight in scenarios where human lives are at risk.

5.6.2. Transparency, Interpretability, and Explainability in EMS AI

Transparency in healthcare AI refers to the way in which stakeholders comprehend an algorithm's behavior and structure. At its best, transparency allows system auditors and physicians to track the conversion of input variables, such as patient symptoms, call metadata, and geolocation factors, into model output [102]. Although they are closely connected, interpretability refers to how well end users can understand the reasoning behind a given choice, especially in complicated or "black-box" models [101]. Furthermore, insufficient interpretability can lead to ambiguous triage recommendations, which makes it challenging for medical regulators and emergency call takers to confirm AI outputs against clinical context and to identify any sources of bias or error [103]. An important finding from this analysis is the widespread absence of explainable AI (XAI) integration in the examined research. As shown in Table 5, the column "XAI" is marked as "No" in the vast majority of cases, indicating that the interpretability of the model was omitted or treated as a secondary consideration. This lack of XAI is particularly challenging in the context of EMS, a high-stakes and safety-critical field where opaque "black-box" decision-making can worsen the effects of inaccurate forecasts, undermining physician trust [104]. One of the most critical gaps identified in this review is the near-absence of explainable AI (XAI) approaches in AI systems for emergency medical services. Although many studies focused on optimizing predictive accuracy and operational efficiency, none systematically integrated explainability frameworks such as SHAP, LIME, counterfactual explanations, or rule-based surrogate models. This absence is not a trivial omission. In high-stakes, time-critical medical decision-making environments, the absence of explanatory power undermines trust, impedes clinical adoption, and poses legal and ethical risks [105]. Future research should focus on incorporating XAI approaches into EMS AI pipelines, especially in large language model-based triage assistants and dispatch optimization systems, to guarantee that the suggestions produced are clinically accurate and interpretable. A framework combining multimodal input analysis with post-hoc explanation methods could bridge this gap and accelerate regulatory approval.

5.7. Data Labeling and Quality Assurance Challenges

Some studies indicate that data quality in emergency medical services (EMS) is insufficient, frequently leaving critical information inaccessible to healthcare professionals and managers. Emergency call audio is frequently recorded under chaotic conditions, where background noise, overlapping speech, and distressed or unclear articulation degrade signal clarity [106]. These factors significantly impact the performance of automatic speech recognition (ASR) systems. Even advanced models, such as Wav2Vec 2.0, have reported word error rates (WERs) exceeding 42% in real-world emergency scenarios, underscoring the difficulty of achieving accurate transcription in these environments [99]. To ensure that AI models for prehospital care deliver on their promise, high-quality labeled data are essential; however, labeling EMS data can be challenging. For machine learning models to be trained on audio call records, transcription and annotation are necessary. This procedure is frequently carried out manually by people, which takes a lot of time and is prone to

errors. For instance, in a Brazilian study on emergency calls [107], researchers used a program called LabelStudio to manually categorize a training dataset of calls made by medical students. After listening to the calls, the students transcribed them and classified each one according to the type of emergency. Although labor-intensive, this type of manual labeling is required since automatic techniques cannot handle the loud, unprocessed audio. It can be challenging to ensure quality in these labels since individuals may disagree on how to classify a call or make transcription errors, particularly if the audio is difficult to hear. Transcripts and labels are often verified by many reviewers or cross-referenced with official incident reports as part of quality assurance. Inaccurate or incorrectly classified data might confuse the AI during training if not carefully validated. In addition, defining consistent labeling is made more difficult by the diversity of emergency data. According to one study, 61% of the terms in a large dataset of emergency call transcripts occurred only once. This suggests that many calls contain unique facts that do not commonly recur, such as names, precise places, or unusual medical or situational phrases. Due to this variability, an annotator may come across unfamiliar situations or words in every call [108]. It can be challenging to come up with a set of categories to categorize these calls (such as "cardiac arrest," "trauma," "fire," "crime," etc.) because actual situations do not often cleanly fall into one category. Label granularity is another problem. For instance, a dispatcher might assign a call "traffic accident," whereas another dispatcher may call it "vehicle entrapment with injuries." Human annotators must be trained and given clear rules in order to label consistently. While some EMS systems employ formal rules (such as medical priority dispatch codes) to standardize the categorization of calls, it is still difficult for both humans and AI to translate the free-form call audio to the appropriate code. The maintenance of model dependability, however, depends critically on making sure that labeled EMS data is verified and validated. In many research procedures, a representative subset of labeled cases is reviewed by domain specialists, such as emergency doctors or expert paramedics, to verify the correctness and address any mistakes made by non-expert annotators, such as medical students or laypeople [25,45]. By identifying incorrect categories or unclear situations, this expert review can assist in improving labeling standards.

6. Conclusions

Recent advancements in artificial intelligence (AI) and machine learning (ML) are progressively being used to address the escalating complexity within emergency medical services (EMS). This article provides a comprehensive review of how machine learning (ML) and deep learning (DL) techniques have been applied throughout the prehospital care continuum to support emergency operations and improve their overall effectiveness. This analysis looked at 58 studies that focused on the use of AI and ML in prehospital emergency medical services and were published between 2018 and the beginning of 2025. The findings highlight important advancements in using AI for critical clinical and administrative functions, such as care coordination, risk assessment, diagnosis, and triage. To ensure an organized synthesis, the selected articles were analyzed and categorized in accordance with the various stages of EMS episodes and clinical subdomains. The examined groups utilize various machine learning (ML) and deep learning (DL) algorithms, as well as hybrid ML-NLP and DL-NLP frameworks, for multiple prehospital applications, including flood response, out-of-hospital cardiac arrest prediction, emergency department (ED) arrival forecasting, emergency severity index calculation, and other critical tasks. The findings highlight notable advancements in the use of AI for critical clinical and operational activities, including diagnosis, risk assessment, triage, and care coordination. Logistic regression, random forest (RF), and support vector machines (SVMs) are some of the most popular traditional ML models because they are easy to implement and provide good

interpretability. However, deep learning methods, particularly LSTM and Transformer-based architectures, have become more popular due to the growing presence of sequential, unstructured, and multimodal features in EMS data. They perform in dynamic EMS situations due to their ability to represent time series and contextual patterns.

Reviewed studies focused on areas of care pathways, triage and prioritization, diagnostics, early warning systems, disaster monitoring, risk assessment, and resource allocation optimization, as well as case studies and applications of LLM-based methods such as GPT-based diagnostic evaluation and emergency detection. Future research should focus on interpretability and explainability to better leverage robust AI systems capable of explaining to domain experts both the result and the process that led to it, and to improve performance in the application domain of EMS. Accordingly, research should focus on improving data quality and integrating LLM and multimodal data. In addition, efforts should be made to design frameworks that guarantee accessibility and interoperability of health data for researchers while preserving patient privacy and data security [109]. Even while prehospital emergency care AI research is advancing quickly, its full practical potential is still not being completely realized. The fragmented and sometimes inaccessible nature of EMS data, especially for high-frequency, multimodal, and real-time streams, is a recurring challenge. Public health organizations, emergency care providers, university researchers, and private AI developers must work together to overcome this restriction. The development of shared platforms to store high-quality, standardized, and interoperable EMS datasets can be facilitated by the establishment of cooperative structures, such as public-private collaborations.

Furthermore, exploring advanced prompting strategies, LLM-based agentic systems [17], and RAG approaches [110] remains crucial, including the incorporation of more detailed contextual information and diverse examples to better guide the LLM's understanding. Another vital area lies in investigating alternative state-of-the-art LLMs and architectural innovations, including those designed for multimodal data processing [25], capable of integrating data from various patient sources, to leverage the rich information contained in the audio recordings of calls and their transcriptions, and the medical regulation file (Dossier de Régulation Médicale—DRM) shared during calls. The development of multimodal datasets combining free-text emergency communication scenarios and ED records with emotional indicators extracted from callers' audio offers new opportunities for a comprehensive understanding of emergency scenarios, and facilitates faster, better-informed decision-making and responses. Emerging research demonstrates that multimodal LLM [93] such as GPT, ClinicalBERT, and Llama can accurately analyze emergency department reports and provide clinical recommendations for patients evaluated in the emergency department, including life-saving interventions, diagnostic examinations, and radiological imaging [111].

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

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article.

Acknowledgments: This work was supported by the GESICA project (GEstion des SItuations de Crises sAnitaires), funded under the Interreg France–Switzerland 2024–2027 cooperation program. This project is led by Marie & Louis Pasteur University (UMLP) and Geneva University Hospitals (HUG).

Conflicts of Interest: Author Thibaut Desmettre is affiliated with the Emergency Department, HUG Genève, Switzerland. The authors of the manuscript, Omar Elfahim, Kokou Laris Edjinedja, Johan Cossus, Mohamed Youssfi, Oussama Barakat, and Thibaut Desmettre, 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.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
ML	Machine Learning
DL	Deep Learning
NLP	Natural Language Processing
EMS	Emergency Medical Services
ED	Emergency Department
LLMs	Large Language Models
RL	Reinforcement Learning
RF	Random Forest
DT	Decision Tree
XGBoost	Extreme Gradient Boosting
GB	Gradient Boosting
LightGBM	Light Gradient Boosting Machine
K-means	K-means Clustering
DBSCAN	Density-Based Spatial Clustering
LR	Logistic Regression
MLP	Multi-Layer Perceptron
ANN	Artificial Neural Network
DNN	Deep Neural Network
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
BiLSTM	Bidirectional LSTM
FCN	Fully Convolutional Network
Transformer	Transformer Architecture
BERT	Bidirectional Encoder Representations from Transformers
GPT-3.5	Generative Pretrained Transformer 3.5
GPT-4	Generative Pretrained Transformer 4
ClinicalBERT	Clinical Bidirectional Encoder Representations from Transformers
Llama	Large Language Model Meta AI
TF-IDF	Term Frequency–Inverse Document Frequency
NBayes	Naive Bayes
FastText	Facebook's Text Classification Model

Appendix A

Table A1. Evaluation metrics for AI Models in prehospital care.

Metric	Formula	Type	Usage in Prehospital AI	Considerations	Ref.
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	Classification	Measures overall correctness of predictions	Not suitable for imbalanced datasets	[2,51]
Weighted accuracy (WA)	$\sum_{i=1}^{C} w_i \cdot \text{accuracy}_i$	Classification	Adjusts accuracy based on class distribution	Reduces bias toward majority classes	[25]
Unweighted accuracy (UWA)	$\frac{1}{C}\sum_{i=1}^{C} \text{accuracy}_i$	Classification	Averages individual class accuracies	More reliable for imbalanced datasets	[90]
Precision	$\frac{TP}{TP+FP}$	Classification	Measures correctness of positive predictions	Important for critical emergency diagnoses	[30,56]
Recall (Sensitivity)	$\frac{TP}{TP+FN}$	Classification	Measures ability to detect actual positives	Crucial for high-risk patient identification	[43,65]
F1 score	$2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision}+\text{Recall}}$	Classification	Balances precision and recall	Used when both false positives and negatives matter	[32,38]
ROC-AUC	Integral of ROC curve	Classification	Evaluates trade-off between TPR and FPR	Higher values indicate better discrimination	[9,57]
RMSE	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$	Regression	Measures deviation of predictions from actual values	Sensitive to large errors	[24]
MAE	$\frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y}_i \rvert$	Regression	Measures average absolute prediction error	Less sensitive to outliers than RMSE	[24]
BLEU score	$\text{BP} \cdot \exp\left(\sum_{n=1}^{4} w_n \log p_n\right)$	LLM	Evaluates quality of generated text translations/reports	Focuses on n-gram precision	[112]
ROUGE-L	F-measure of LCS	LLM	Assesses summary quality in emergency reports	Measures longest common subsequence with references	[112]
Hallucination rate	$\frac{\text{Unsupported claims}}{\text{Total claims}}$	LLM	Quantifies fictitious information generation	Vital for reliable triage recommendations	[112]

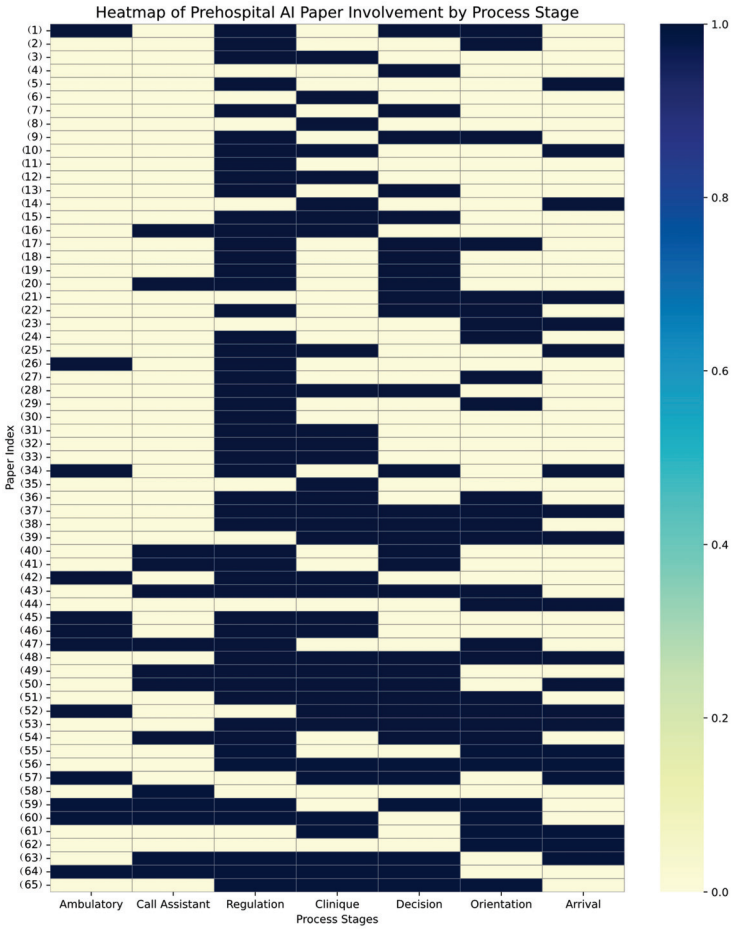


Figure A1. Visualization of prehospital-AI paper interest by process stage.

References

1. Gil-Jardiné, C.; Chenais, G.; Pradeau, C.; Tentillier, E.; Revel, P.; Combes, X.; Galinski, M.; Tellier, E.; Lagarde, E. Trends in reasons for emergency calls during the COVID-19 crisis in the department of Gironde, France using artificial neural network for natural language classification. *Scand. J. Trauma Resusc. Emerg. Med.* **2021**, *29*, 1–9. [CrossRef]
2. Linardos, V.; Drakaki, M.; Tzionas, P.; Karnavas, Y.L. Machine learning in disaster management: Recent developments in methods and applications. *Mach. Learn. Knowl. Extr.* **2022**, *4*, 446–473. [CrossRef]
3. Sterling, N.W.; Patzer, R.E.; Di, M.; Schrager, J.D. Prediction of emergency department patient disposition based on natural language processing of triage notes. *Int. J. Med Inform.* **2019**, *129*, 184–188. [CrossRef] [PubMed]
4. Harford, S.; Del Rios, M.; Heinert, S.; Weber, J.; Markul, E.; Tataris, K.; Campbell, T.; Vanden Hoek, T.; Darabi, H. A machine learning approach for modeling decisions in the out of hospital cardiac arrest care workflow. *BMC Med Inform. Decis. Mak.* **2022**, *22*, 21. [CrossRef] [PubMed]
5. Ahmed, Z.; Mohamed, K.; Zeeshan, S.; Dong, X. Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine. *Database* **2020**, *2020*, baaa010. [CrossRef]
6. Alom, M.Z.; Taha, T.M.; Yakopcic, C.; Westberg, S.; Sidike, P.; Nasrin, M.S.; Hasan, M.; Van Essen, B.C.; Awwal, A.A.; Asari, V.K. A state-of-the-art survey on deep learning theory and architectures. *Electronics* **2019**, *8*, 292. [CrossRef]
7. Willemink, M.J.; Koszek, W.A.; Hardell, C.; Wu, J.; Fleischmann, D.; Harvey, H.; Folio, L.R.; Summers, R.M.; Rubin, D.L.; Lungren, M.P. Preparing medical imaging data for machine learning. *Radiology* **2020**, *295*, 4–15. [CrossRef]
8. Gulshan, V.; Peng, L.; Coram, M.; Stumpe, M.C.; Wu, D.; Narayanaswamy, A.; Venugopalan, S.; Widner, K.; Madams, T.; Cuadros, J.; et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* **2016**, *316*, 2402–2410. [CrossRef]
9. Choi, E.; Schuetz, A.; Stewart, W.F.; Sun, J. Using recurrent neural network models for early detection of heart failure onset. *J. Am. Med Inform. Assoc.* **2017**, *24*, 361–370. [CrossRef]
10. Gong, F.; Wang, M.; Wang, H.; Wang, S.; Liu, M. SMR: Medical knowledge graph embedding for safe medicine recommendation. *Big Data Res.* **2021**, *23*, 100174. [CrossRef]
11. Kang, D.Y.; Cho, K.J.; Kwon, O.; Kwon, J.m.; Jeon, K.H.; Park, H.; Lee, Y.; Park, J.; Oh, B.H. Artificial intelligence algorithm to predict the need for critical care in prehospital emergency medical services. *Scand. J. Trauma Resusc. Emerg. Med.* **2020**, *28*, 17. [CrossRef]
12. Rafi, S.; Gangloff, C.; Paulhet, E.; Grimault, O.; Soulat, L.; Bouzillé, G.; Cuggia, M. Out-of-hospital cardiac arrest detection by machine learning based on the phonetic characteristics of the caller's voice. In *Challenges of Trustable AI and Added-Value on Health*; IOS Press: Amsterdam, The Netherlands, 2022; pp. 445–449.
13. Stewart, J.; Lu, J.; Goudie, A.; Arendts, G.; Meka, S.A.; Freeman, S.; Walker, K.; Sprivulis, P.; Sanfilippo, F.; Bennamoun, M.; et al. Applications of natural language processing at emergency department triage: A narrative review. *PLoS ONE* **2023**, *18*, e0279953. [CrossRef]
14. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
15. Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. Gpt-4 technical report. *arXiv* **2023**, arXiv:2303.08774. [CrossRef]
16. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. Llama: Open and efficient foundation language models. *arXiv* **2023**, arXiv:2302.13971. [CrossRef]
17. Qiu, J.; Lam, K.; Li, G.; Acharya, A.; Wong, T.Y.; Darzi, A.; Yuan, W.; Topol, E.J. LLM-based agentic systems in medicine and healthcare. *Nat. Mach. Intell.* **2024**, *6*, 1418–1420. [CrossRef]
18. Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; DasSarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv* **2022**, arXiv:2204.05862. [CrossRef]
19. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* **2021**, *372*, n71. [CrossRef] [PubMed]
20. Shi, J.; Ye, M.; Chen, H.; Lu, Y.; Tan, Z.; Fan, Z.; Zhao, J. Enhancing efficiency and capacity of telehealth services with intelligent triage: A bidirectional LSTM neural network model employing character embedding. *BMC Med Inform. Decis. Mak.* **2023**, *23*, 269. [CrossRef] [PubMed]
21. Ceklic, E.; Ball, S.; Finn, J.; Brown, E.; Brink, D.; Bailey, P.; Whiteside, A.; Brits, R.; Tohira, H. Ambulance dispatch prioritisation for traffic crashes using machine learning: A natural language approach. *Int. J. Med Inform.* **2022**, *168*, 104886. [CrossRef]
22. Inokuchi, R.; Iwagami, M.; Sun, Y.; Sakamoto, A.; Tamiya, N. Machine learning models predicting undertriage in telephone triage. *Ann. Med.* **2022**, *54*, 2989–2996. [CrossRef] [PubMed]

23. Rojas, A.; Calcín, K.; Casta neda, P. A Whisper and BETO-Bert Based Web Application for Classification of Emergency Calls. In Proceedings of the 2025 8th International Conference on Information and Computer Technologies (ICICT), Honolulu, HI, USA, 14–16 March 2025; pp. 236–241.
24. Brossard, C.; Goetz, C.; Catoire, P.; Cipolat, L.; Guyeux, C.; Gil Jardine, C.; Akplogan, M.; Abensur Vuillaume, L. Predicting emergency department admissions using a machine-learning algorithm: A proof of concept with retrospective study. *BMC Emerg. Med.* **2025**, *25*, 3. [CrossRef]
25. Ferri, P.; Sáez, C.; Félix-De Castro, A.; Juan-Albarracín, J.; Blanes-Selva, V.; Sánchez-Cuesta, P.; García-Gómez, J.M. Deep ensemble multitask classification of emergency medical call incidents combining multimodal data improves emergency medical dispatch. *Artif. Intell. Med.* **2021**, *117*, 102088. [CrossRef]
26. Abi Kanaan, M.; Couchot, J.F.; Guyeux, C.; Laiymani, D.; Atechian, T.; Darazi, R. Combining a multi-feature neural network with multi-task learning for emergency calls severity prediction. *Array* **2024**, *21*, 100333. [CrossRef]
27. Chin, K.C.; Hsieh, T.C.; Chiang, W.C.; Chien, Y.C.; Sun, J.T.; Lin, H.Y.; Hsieh, M.J.; Yang, C.W.; Chen, A.Y.; Ma, M.H.M. Early recognition of a caller's emotion in out-of-hospital cardiac arrest dispatching: An artificial intelligence approach. *Resuscitation* **2021**, *167*, 144–150. [CrossRef] [PubMed]
28. Mutegeki, H.; Nahabwe, A.; Nakatumba-Nabende, J.; Marvin, G. Interpretable machine learning-based triage for decision support in emergency care. In Proceedings of the 2023 7th International Conference on Trends in Electronics and Informatics (ICOEI), Tamil Nadu, India, 27–29 April 2023; pp. 983–990.
29. Alkhodair, S.A.; Altwaijri, N.; Albarrak, A.I. Identifying Preventable Emergency Admissions in Hospitals Using Machine Learning. In *Telehealth Ecosystems in Practice*; IOS Press: Amsterdam, The Netherlands, 2023; pp. 95–96.
30. Kim, J.H.; Kim, B.; Kim, M.J.; Hyun, H.; Kim, H.C.; Chang, H.J. Prediction of inappropriate pre-hospital transfer of patients with suspected cardiovascular emergency diseases using machine learning: A retrospective observational study. *BMC Med Inform. Decis. Mak.* **2023**, *23*, 56. [CrossRef]
31. Torres, N.; Trujillo, L.; Maldonado, Y.; Vera, C. Correction of the travel time estimation for ambulances of the red cross Tijuana using machine learning. *Comput. Biol. Med.* **2021**, *137*, 104798. [CrossRef]
32. Yang, L.; Liu, Q.; Zhao, Q.; Zhu, X.; Wang, L. Machine learning is a valid method for predicting prehospital delay after acute ischemic stroke. *Brain Behav.* **2020**, *10*, e01794. [CrossRef] [PubMed]
33. Lin, A.X.; Ho, A.F.W.; Cheong, K.H.; Li, Z.; Cai, W.; Chee, M.L.; Ng, Y.Y.; Xiao, X.; Ong, M.E.H. Leveraging machine learning techniques and engineering of multi-nature features for national daily regional ambulance demand prediction. *Int. J. Environ. Res. Public Health* **2020**, *17*, 4179. [CrossRef] [PubMed]
34. Spangler, D.; Hermansson, T.; Smekal, D.; Blomberg, H. A validation of machine learning-based risk scores in the prehospital setting. *PLoS ONE* **2019**, *14*, e0226518. [CrossRef]
35. Redfield, C.; Tlimat, A.; Halpern, Y.; Schoenfeld, D.W.; Ullman, E.; Sontag, D.A.; Nathanson, L.A.; Horng, S. Derivation and validation of a machine learning record linkage algorithm between emergency medical services and the emergency department. *J. Am. Med Inform. Assoc.* **2020**, *27*, 147–153. [CrossRef]
36. Arnaud, E.; Elbattah, M.; Moreno-Sánchez, P.A.; Dequen, G.; Ghazali, D.A. Explainable NLP model for predicting patient admissions at emergency department using triage notes. In Proceedings of the 2023 IEEE International Conference on Big Data (BigData), Sorrento, Italy, 15–18 December 2023; pp. 4843–4847.
37. Chen, A.Y.; Lu, T.Y.; Ma, M.H.M.; Sun, W.Z. Demand forecast using data analytics for the preallocation of ambulances. *IEEE J. Biomed. Health Inform.* **2015**, *20*, 1178–1187. [CrossRef]
38. Moon, H.J.; Shin, Y.J.; Cho, Y.S. Identification of out-of-hospital cardiac arrest clusters using unsupervised learning. *Am. J. Emerg. Med.* **2022**, *62*, 41–48. [CrossRef]
39. Hirano, Y.; Kondo, Y.; Sueyoshi, K.; Okamoto, K.; Tanaka, H. Early outcome prediction for out-of-hospital cardiac arrest with initial shockable rhythm using machine learning models. *Resuscitation* **2021**, *158*, 49–56. [CrossRef]
40. Al-Dury, N.; Ravn-Fischer, A.; Hollenberg, J.; Israelsson, J.; Nordberg, P.; Strömsöe, A.; Axelsson, C.; Herlitz, J.; Rawshani, A. Identifying the relative importance of predictors of survival in out of hospital cardiac arrest: A machine learning study. *Scand. J. Trauma Resusc. Emerg. Med.* **2020**, *28*, 1–8. [CrossRef] [PubMed]
41. Bakidou, A.; Caragounis, E.C.; Andersson Hagiwara, M.; Jonsson, A.; Sjöqvist, B.A.; Candefjord, S. On Scene Injury Severity Prediction (OSISP) model for trauma developed using the Swedish Trauma Registry. *BMC Med Inform. Decis. Mak.* **2023**, *23*, 206. [CrossRef] [PubMed]
42. Wang, J.; Zhang, J.; Gong, X.; Zhang, W.; Zhou, Y.; Lou, M. Prediction of large vessel occlusion for ischaemic stroke by using the machine learning model random forests. *Stroke Vasc. Neurol.* **2022**, *7*, e001096. [CrossRef] [PubMed]
43. Strum, R.P.; Mowbray, F.I.; Zargoush, M.; Jones, A.P. Prehospital prediction of hospital admission for emergent acuity patients transported by paramedics: A population-based cohort study using machine learning. *PLoS ONE* **2023**, *18*, e0289429. [CrossRef] [PubMed]

44. Ferri, P.; Lomonaco, V.; Passaro, L.C.; Félix-De Castro, A.; Sánchez-Cuesta, P.; Sáez, C.; García-Gómez, J.M. Deep continual learning for medical call incidents text classification under the presence of dataset shifts. *Comput. Biol. Med.* **2024**, *175*, 108548. [CrossRef]
45. Paulin, J.; Reunamo, A.; Kurola, J.; Moen, H.; Salanterä, S.; Riihimäki, H.; Vesanen, T.; Koivisto, M.; Iirola, T. Using machine learning to predict subsequent events after EMS non-conveyance decisions. *BMC Med Inform. Decis. Mak.* **2022**, *22*, 166. [CrossRef]
46. Kitano, S.; Ogawa, K.; Igarashi, Y.; Nishimura, K.; Osawa, S.; Suzuki, K.; Fujimoto, K.; Harada, S.; Narikawa, K.; Tagami, T.; et al. Development of a Machine Learning Model to Predict Cardiac Arrest during Transport of Trauma Patients. *J. Nippon Med Sch.* **2023**, *90*, 186–193. [CrossRef]
47. Carrell, D.S.; Gruber, S.; Floyd, J.S.; Bann, M.A.; Cushing-Haugen, K.L.; Johnson, R.L.; Graham, V.; Cronkite, D.J.; Hazlehurst, B.L.; Felcher, A.H.; et al. Improving methods of identifying anaphylaxis for medical product safety surveillance using natural language processing and machine learning. *Am. J. Epidemiol.* **2023**, *192*, 283–295. [CrossRef] [PubMed]
48. Zhang, Z.; Zhou, D.; Zhang, J.; Xu, Y.; Lin, G.; Jin, B.; Liang, Y.; Geng, Y.; Zhang, S. Multilayer perceptron-based prediction of stroke mimics in prehospital triage. *Sci. Rep.* **2022**, *12*, 17994. [CrossRef]
49. Kawai, Y.; Okuda, H.; Kinoshita, A.; Yamamoto, K.; Miyazaki, K.; Takano, K.; Asai, H.; Urisono, Y.; Fukushima, H. Visual assessment of interactions among resuscitation activity factors in out-of-hospital cardiopulmonary arrest using a machine learning model. *PLoS ONE* **2022**, *17*, e0273787. [CrossRef]
50. Takeda, M.; Oami, T.; Hayashi, Y.; Shimada, T.; Hattori, N.; Tateishi, K.; Miura, R.E.; Yamao, Y.; Abe, R.; Kobayashi, Y.; et al. Prehospital diagnostic algorithm for acute coronary syndrome using machine learning: A prospective observational study. *Sci. Rep.* **2022**, *12*, 14593. [CrossRef]
51. Duceau, B.; Alsac, J.; Bellenfant, F.; Mailloux, A.; Champigneulle, B.; Favé, G.; Neuschwander, A.; El Batti, S.; Cholley, B.; Achouh, P.; et al. Prehospital triage of acute aortic syndrome using a machine learning algorithm. *J. Br. Surg.* **2020**, *107*, 995–1003. [CrossRef] [PubMed]
52. Chin, K.C.; Cheng, Y.C.; Sun, J.T.; Ou, C.Y.; Hu, C.H.; Tsai, M.C.; Ma, M.H.M.; Chiang, W.C.; Chen, A.Y. Machine Learning–Based Text Analysis to Predict Severely Injured Patients in Emergency Medical Dispatch: Model Development and Validation. *J. Med Internet Res.* **2022**, *24*, e30210. [CrossRef]
53. Abe, D.; Inaji, M.; Hase, T.; Takahashi, S.; Sakai, R.; Ayabe, F.; Tanaka, Y.; Otomo, Y.; Maehara, T. A prehospital triage system to detect traumatic intracranial hemorrhage using machine learning algorithms. *JAMA Netw. Open* **2022**, *5*, e2216393. [CrossRef]
54. Hayashi, Y.; Shimada, T.; Hattori, N.; Shimazui, T.; Yoshida, Y.; Miura, R.E.; Yamao, Y.; Abe, R.; Kobayashi, E.; Iwadate, Y.; et al. A prehospital diagnostic algorithm for strokes using machine learning: A prospective observational study. *Sci. Rep.* **2021**, *11*, 20519. [CrossRef]
55. Uchida, K.; Kouno, J.; Yoshimura, S.; Kinjo, N.; Sakakibara, F.; Araki, H.; Morimoto, T. Development of machine learning models to predict probabilities and types of stroke at prehospital stage: The Japan urgent stroke triage score using machine learning (JUST-ML). *Transl. Stroke Res.* **2022**, *13*, 370–381. [CrossRef]
56. Blomberg, S.N.; Folke, F.; Ersbøll, A.K.; Christensen, H.C.; Torp-Pedersen, C.; Sayre, M.R.; Counts, C.R.; Lippert, F.K. Machine learning as a supportive tool to recognize cardiac arrest in emergency calls. *Resuscitation* **2019**, *138*, 322–329. [CrossRef] [PubMed]
57. Chu, J.; Leung, K.B.; Snobelen, P.; Nevils, G.; Drennan, I.R.; Cheskes, S.; Chan, T.C. Machine learning-based dispatch of drone-delivered defibrillators for out-of-hospital cardiac arrest. *Resuscitation* **2021**, *162*, 120–127. [CrossRef]
58. Zhang, X.; Zhang, H.; Sheng, L.; Tian, F. DL-PER: Deep learning model for Chinese prehospital emergency record classification. *IEEE Access* **2022**, *10*, 64638–64649. [CrossRef]
59. Peng, C.; Peng, L.; Yang, F.; Yu, H.; Chen, Q.; Guo, Y.; Xu, S.; Jin, Z. The prediction of the survival in patients with severe trauma during prehospital care: Analyses based on NTDB database. *Eur. J. Trauma Emerg. Surg.* **2024**, *50*, 1599–1609. [CrossRef] [PubMed]
60. Wenstrup, J.; Havtorn, J.D.; Borgholt, L.; Blomberg, S.N.; Maaloe, L.; Sayre, M.R.; Christensen, H.; Kruuse, C.; Christensen, H.C. A retrospective study on machine learning-assisted stroke recognition for medical helpline calls. *NPJ Digit. Med.* **2023**, *6*, 235. [CrossRef] [PubMed]
61. Choi, A.; Kim, M.J.; Sung, J.M.; Kim, S.; Lee, J.; Hyun, H.; Kim, H.C.; Kim, J.H.; Chang, H.J. Development of prediction models for acute myocardial infarction at prehospital stage with machine learning based on a nationwide database. *J. Cardiovasc. Dev. Dis.* **2022**, *9*, 430. [CrossRef]
62. Kim, D.; You, S.; So, S.; Lee, J.; Yook, S.; Jang, D.P.; Kim, I.Y.; Park, E.; Cho, K.; Cha, W.C.; et al. A data-driven artificial intelligence model for remote triage in the prehospital environment. *PLoS ONE* **2018**, *13*, e0206006. [CrossRef]
63. Anthony, T.; Mishra, A.K.; Stassen, W.; Son, J. The feasibility of using machine learning to classify calls to South African emergency dispatch centres according to prehospital diagnosis, by utilising caller descriptions of the incident. *Healthcare* **2021**, *9*, 1107. [CrossRef] [PubMed]
64. Tignanelli, C.J.; Silverman, G.M.; Lindemann, E.A.; Trembley, A.L.; Gipson, J.C.; Beilman, G.; Lyng, J.W.; Finzel, R.; McEwan, R.; Knoll, B.C.; et al. Natural language processing of prehospital emergency medical services trauma records allows for automated characterization of treatment appropriateness. *J. Trauma Acute Care Surg.* **2020**, *88*, 607–614. [CrossRef]

65. Harford, S.; Darabi, H.; Del Rios, M.; Majumdar, S.; Karim, F.; Vanden Hoek, T.; Erwin, K.; Watson, D.P. A machine learning based model for Out of Hospital cardiac arrest outcome classification and sensitivity analysis. *Resuscitation* **2019**, *138*, 134–140. [CrossRef]
66. Forberg, J.L.; Khoshnood, A.; Green, M.; Ohlsson, M.; Björk, J.; Jovinge, S.; Edenbrandt, L.; Ekelund, U. An artificial neural network to safely reduce the number of ambulance ECGs transmitted for physician assessment in a system with prehospital detection of ST elevation myocardial infarction. *Scand. J. Trauma Resusc. Emerg. Med.* **2012**, *20*, 1–9. [CrossRef]
67. Alonso, E.; Irusta, U.; Aramendi, E.; Daya, M.R. A machine learning framework for pulse detection during out-of-hospital cardiac arrest. *IEEE Access* **2020**, *8*, 161031–161041. [CrossRef]
68. Silverman, G.M.; Lindemann, E.A.; Rajamani, G.; Finzel, R.L.; McEwan, R.; Knoll, B.C.; Pakhomov, S.; Melton, G.B.; Tignanelli, C.J. Named entity recognition in prehospital trauma care. In *MEDINFO 2019: Health and Wellbeing e-Networks for All*; IOS Press: Amsterdam, The Netherlands, 2019; pp. 1586–1587.
69. Eastwood, K.W.; May, R.; Andreou, P.; Abidi, S.; Abidi, S.S.R.; Loubani, O.M. Needs and expectations for artificial intelligence in emergency medicine according to Canadian physicians. *BMC Health Serv. Res.* **2023**, *23*, 798. [CrossRef]
70. Mueller, S.; Murray, M.; Schnipper, J.; Goralnick, E. An initiative to improve advanced notification of inter-hospital transfers. *Healthcare* **2020**, *8*, 100423. [CrossRef]
71. Chu, Z.; Wang, Y.; Zhu, F.; Yu, L.; Li, L.; Gu, J. Professional Agents - Evolving Large Language Models into Autonomous Experts with Human-Level Competencies. *arXiv* **2024**, arXiv:2402.03628.
72. Tang, X.; Zou, A.; Zhang, Z.; Zhao, Y.; Zhang, X.; Cohan, A.; Gerstein, M.B. MedAgents: Large Language Models as Collaborators for Zero-shot Medical Reasoning. *arXiv* **2023**, arXiv:2311.10537.
73. Arslan, B.; Nuhoglu, C.; Satici, M.; Altinbilek, E. Evaluating LLM-based generative AI tools in emergency triage: A comparative study of ChatGPT Plus, Copilot Pro, and triage nurses. *Am. J. Emerg. Med.* **2025**, *89*, 174–181. [CrossRef] [PubMed]
74. Li, Y.; Li, Z.; Zhang, K.; Dan, R.; Jiang, S.; Zhang, Y. ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge. *Cureus* **2023**, *13*, e40895. [CrossRef]
75. Guo, T.; Chen, X.; Wang, Y.; Chang, R.; Pei, S.; Chawla, N.; Wiest, O.; Zhang, X. Large Language Model based Multi-Agents: A Survey of Progress and Challenges. *arXiv* **2024**, arXiv:2402.01680. [CrossRef]
76. Haim, G.B.; Braun, A.; Eden, H.; Burshtein, L.; Barash, Y.; Irony, A.; Klang, E. AI in the ED: Assessing the efficacy of GPT models vs. physicians in medical score calculation. *Am. J. Emerg. Med.* **2024**, *79*, 161–166. [CrossRef]
77. Lorenzoni, G.; Gregori, D.; Bressan, S.; Ocagli, H.; Azzolina, D.; Da Dalt, L.; Berchialla, P. Use of a large language model to identify and classify injuries with free-text emergency department data. *JAMA Netw. Open* **2024**, *7*, e2413208. [CrossRef] [PubMed]
78. Ventura, C.A.I.; Denton, E.E. Artificial intelligence chatbots and emergency medical services: Perspectives on the implications of generative AI in prehospital care. In *Open Access Emergency Medicine*; Taylor & Francis: Abingdon, UK, 2023, pp. 289–292.
79. Akaybicen, F.; Cummings, A.; Iwuagwu, L.; Zhang, X.; Adewuyi, M. A Machine Learning Approach for Emergency Detection in Medical Scenarios Using Large Language Models. *arXiv* **2024**, arXiv:2412.16341. [CrossRef]
80. Shekhar, A.C.; Kimbrell, J.; Saharan, A.; Stebel, J.; Ashley, E.; Abbott, E.E. Use of a large language model (LLM) for ambulance dispatch and triage. *Am. J. Emerg. Med.* **2025**, *89*, 27–29. [CrossRef] [PubMed]
81. Cui, W.; Kawamoto, K.; Morgan, K.; Finkelstein, J. Reducing Diagnostic Uncertainty in Emergency Departments: The Role of Large Language Models in Age-Specific Diagnostics. In Proceedings of the 2024 IEEE 12th International Conference on Healthcare Informatics (ICHI), Orlando, FL, USA, 3–6 June 2024; pp. 525–527.
82. Shah-Mohammadi, F.; Finkelstein, J. Accuracy evaluation of GPT-assisted differential diagnosis in emergency department. *Diagnostics* **2024**, *14*, 1779. [CrossRef]
83. Colakca, C.; Ergın, M.; Ozensoy, H.S.; Sener, A.; Guru, S.; Ozhasenekler, A. Emergency department triaging using ChatGPT based on emergency severity index principles: A cross-sectional study. *Sci. Rep.* **2024**, *14*, 22106. [CrossRef] [PubMed]
84. Williams, C.Y.; Zack, T.; Miao, B.Y.; Sushil, M.; Wang, M.; Kornblith, A.E.; Butte, A.J. Use of a large language model to assess clinical acuity of adults in the emergency department. *JAMA Netw. Open* **2024**, *7*, E248895. [CrossRef]
85. Montagna, S.; Ferretti, S.; Klopfenstein, L.C.; Ungolo, M.; Pengo, M.F.; Aguzzi, G.; Magnini, M. Privacy-preserving LLM-based chatbots for hypertensive patient self-management. *Smart Health* **2025**, *36*, 100552. [CrossRef]
86. Lin, C.; Kuo, C.F. Roles and potential of large language models in healthcare: A comprehensive review. *Biomed. J.* **2025**, 100868. [CrossRef]
87. Amugongo, L.M.; Mascheroni, P.; Brooks, S.; Doering, S.; Seidel, J. Retrieval augmented generation for large language models in healthcare: A systematic review. *PLoS Digit. Health* **2025**, *4*, e0000877. [CrossRef]
88. Karunanayake, N. Next-generation agentic AI for transforming healthcare. *Inform. Health* **2025**, *2*, 73–83. [CrossRef]
89. Wang, J.; Shi, E.; Yu, S.; Wu, Z.; Ma, C.; Dai, H.; Yang, Q.; Kang, Y.; Wu, J.; Hu, H.; et al. Prompt engineering for healthcare: Methodologies and applications. *arXiv* **2023**, arXiv:2304.14670.

90. Deschamps-Berger, T.; Lamel, L.; Devillers, L. End-to-end speech emotion recognition: Challenges of real-life emergency call centers data recordings. In Proceedings of the 2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII), Virtual, 28 September–1 October 2021; pp. 1–8.
91. Brant, E.B.; Kennedy, J.N.; King, A.J.; Gerstley, L.D.; Mishra, P.; Schlessinger, D.; Shalaby, J.; Escobar, G.J.; Angus, D.C.; Seymour, C.W.; et al. Developing a shared sepsis data infrastructure: A systematic review and concept map to FHIR. *NPJ Digit. Med.* **2022**, *5*, 44. [CrossRef]
92. Miller, M.; Jorm, L.; Partyka, C.; Burns, B.; Habig, K.; Oh, C.; Immens, S.; Ballard, N.; Gallego, B. Identifying prehospital trauma patients from ambulance patient care records; comparing two methods using linked data in New South Wales, Australia. *Injury* **2024**, *55*, 111570. [CrossRef] [PubMed]
93. Pasquadibisceglie, V.; Appice, A.; Malerba, D.; Fiameni, G. Leveraging a Large Language Model (LLM) to Predict Hospital Admissions of Emergency Department Patients. *Expert Syst. Appl.* **2025**, *128224*, 128224. [CrossRef]
94. Douglass, M.; Clifford, G.D.; Reisner, A.; Moody, G.B.; Mark, R.G. Computer-assisted de-identification of free text in the MIMIC II database. In Proceedings of the Computers in Cardiology, Chicago, IL, USA, 19–22 September 2004; pp. 341–344.
95. Dyda, A.; Purcell, M.; Curtis, S.; Field, E.; Pillai, P.; Ricardo, K.; Weng, H.; Moore, J.C.; Hewett, M.; Williams, G.; et al. Differential privacy for public health data: An innovative tool to optimize information sharing while protecting data confidentiality. *Patterns* **2021**, *2*, 100366. [CrossRef] [PubMed]
96. Voigt, P.; Von dem Bussche, A. The eu general data protection regulation (gdpr). In *A Practice Guide*, 1st ed.; Springer International Publishing: Cham, Switzerland, 2017; Volume 10, pp. 10–5555.
97. Fix, J.; Ising, A.I.; Proescholdbell, S.K.; Falls, D.M.; Wolff, C.S.; Fernandez, A.R.; Waller, A.E. Linking emergency medical services and emergency department data to improve overdose surveillance in North Carolina. *Public Health Rep.* **2021**, *136*, 54S–61S. [CrossRef]
98. Newgard, C.D. Validation of probabilistic linkage to match de-identified ambulance records to a state trauma registry. *Acad. Emerg. Med.* **2006**, *13*, 69–75. [CrossRef]
99. Weerasinghe, K.; Janapati, S.; Ge, X.; Kim, S.; Iyer, S.; Stankovic, J.A.; Alemzadeh, H. Real-Time Multimodal Cognitive Assistant for Emergency Medical Services. *arXiv* **2024**, arXiv:2403.06734. [CrossRef]
100. Pandey, D.; Jahanabadi, H.; D'Arcy, J.; Doherty, S.; Vo, H.; Jones, D.; Bellomo, R. Early prediction of intensive care unit admission in emergency department patients using machine learning. *Aust. Crit. Care* **2025**, *38*, 101143. [CrossRef]
101. Amann, J.; Blasimme, A.; Vayena, E.; Frey, D.; Madai, V.I.; Consortium, P. Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. *BMC Med Inform. Decis. Mak.* **2020**, *20*, 310. [CrossRef] [PubMed]
102. Bernal, J.; Mazo, C. Transparency of artificial intelligence in healthcare: Insights from professionals in computing and healthcare worldwide. *Appl. Sci.* **2022**, *12*, 10228. [CrossRef]
103. Rajpurkar, P.; Chen, E.; Banerjee, O.; Topol, E.J. AI in health and medicine. *Nat. Med.* **2022**, *28*, 31–38. [CrossRef]
104. Chaddad, A.; Peng, J.; Xu, J.; Bouridane, A. Survey of explainable AI techniques in healthcare. *Sensors* **2023**, *23*, 634. [CrossRef]
105. Adadi, A.; Berrada, M. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access* **2018**, *6*, 52138–52160. [CrossRef]
106. Mashoufi, M.; Ayatollahi, H.; Khorasani-Zavareh, D.; Talebi Azad Boni, T. Data quality assessment in emergency medical services: An objective approach. *BMC Emerg. Med.* **2023**, *23*, 10. [CrossRef] [PubMed]
107. Costa, D.B.; Pinna, F.C.d.A.; Joiner, A.P.; Rice, B.; Souza, J.V.P.d.; Gabella, J.L.; Andrade, L.; Vissoci, J.R.N.; Néto, J.C. AI-based approach for transcribing and classifying unstructured emergency call data: A methodological proposal. *PLoS Digit. Health* **2023**, *2*, e0000406. [CrossRef]
108. Valizada, A.; Akhundova, N.; Rustamov, S. Development of speech recognition systems in emergency call centers. *Symmetry* **2021**, *13*, 634. [CrossRef]
109. Okada, Y.; Ning, Y.; Ong, M.E.H. Explainable artificial intelligence in emergency medicine: An overview. *Clin. Exp. Emerg. Med.* **2023**, *10*, 354. [CrossRef] [PubMed]
110. Ng, K.K.Y.; Matsuba, I.; Zhang, P.C. RAG in health care: A novel framework for improving communication and decision-making by addressing LLM limitations. *NEJM AI* **2025**, *2*, AIra2400380. [CrossRef]
111. Huang, T.Y.; Chong, C.F.; Lin, H.Y.; Chen, T.Y.; Chang, Y.C.; Lin, M.C. A pre-trained language model for emergency department intervention prediction using routine physiological data and clinical narratives. *Int. J. Med Inform.* **2024**, *191*, 105564. [CrossRef] [PubMed]
112. Hartman, V.; Zhang, X.; Poddar, R.; McCarty, M.; Fortenko, A.; Sholle, E.; Sharma, R.; Campion, T.; Steel, P.A. Developing and evaluating large language model–generated emergency medicine handoff notes. *JAMA Netw. Open* **2024**, *7*, e2448723. [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.

AI

MDPI

Article

Multi-Objective Optimization Model for Emergency Evacuation Based on Adaptive Ant Colony Algorithm

Jiacheng Yuan and Baiqing Sun *

School of Management, Harbin Institute of Technology, Harbin 150001, China; jiacheng_y2025@163.com
* Correspondence: baiqingsun@hit.edu.cn

Abstract

Evacuation in public places under emergency situations represents a significant area of management research. With the rapid development of the railway industry, the evacuation of railway stations has gradually attracted attention. This article employs the minimization of congestion degree and total evacuation time as primary objectives. In addition, the psychological behavior of individuals and the impact of congestion are sufficiently considered. Moreover, an adaptive Cauchy mutation operator is adopted for flexible population diversity. As a result, a multi-objective optimization model for the evacuation paths is established, with an improved adaptive quantum ant colony algorithm, and a comparison between the model based on adaptive quantum ant colony algorithm and the traditional ant colony model is made.

Keywords: emergency management; ant colony optimization; optimization

1. Introduction

In the contemporary era, emergencies commonly occur in public space. Emergency management has garnered significant interest from a diverse array of individuals and sectors. As a nexus of urban transportation, the railway station plays a pivotal role in fostering the economic growth of a vicinity. It becomes a comprehensive transportation hub, integrating transportation, interchange, and service functions.

Consequently, the evacuation of a considerable number of passengers at railway stations during an event is a challenging task, particularly given the potential for injury. With the expansion of railway networks in various countries, it is imperative to address the issue of emergency evacuation at railway stations. To this end, related research is trying to determine the most efficient and orderly scheme for evacuating passengers and staff, which includes reducing the number of casualties and minimizing property losses. Furthermore, such research will contribute to enhancing the risk response capability of railroad transportation sectors.

The evacuation process in an emergency is influenced by a multitude of factors, including the cluster effect of participants, the operational level of management, the secondary impact of emergencies, the layout of building facilities, etc. This complexity and variability introduce significant uncertainty to the decision-making related to evacuation programs [1,2]. Therefore, scholars in this field typically construct simulation and mathematical models for evacuation experiments. Moreover, in order to enhance the realism and scientific rigor of the evacuation model, scholars have introduced a variety of rules and constraints into the model, with the aim of developing a more reasonable and effective scheme.

In regard to simulation experiments, Altshuler et al. [3] employed ants to simulate herd behavior in the model with the degree of panic as a variable. Their findings corroborated the hypothesis that panic is susceptible to diffusion under insufficient information. This model provides a case of considering the individual behaviour.

For the timeliness of the model, Qian et al. [4] developed an interval parameter fuzzy evacuation management model to facilitate uncertain circumstances. Cova and Johnson [5] developed a dynamic network evacuation model that resolves conflicts of channel intersections in a static network. Li Yanfeng et al. [6] proposed a scheduling mechanism for updating routes with instant traffic information.

Regarding veracity and extension of the model, Zhou et al. [7] employed a fuzzy logic model to investigate the emergency evacuation under the terrorist attacks. Ma [8] developed a multi-objective emergency material transfer path optimization model under the user-balanced traffic assignment for the post-earthquake situation. Huang et al. [9] developed a dynamic switching model to address the uncertainty associated with initial conditions and to obtain global evacuation paths in a dynamic environment. In a study conducted by Li and Zhao [10], the advantages of crowd clustering in emergency evacuation plans was examined, particularly focus on the situation with capacity limitations. In conclusion, the simulation model can be extended as an integrated system, which typically incorporates factors such as building facility layouts, individual behaviors, and path optimization algorithms [11–16].

The application of intelligent algorithms based on soft computing offers a novel avenue for exploring solutions to evacuation-related problems. The most commonly used intelligent algorithms for solving evacuation path optimization problems include particle swarm algorithms, genetic algorithms, and ant colony algorithms, etc. Zou et al. [17] propose a multi-objective ant colony optimization model to address the problem of large-scale emergency evacuation management. Zhang Jianghua et al. [18] developed a multi-source evacuation model for addressing the priority order problem with multiple emergency source points and considering the capacity limitations of channels. They also heuristically devised an algorithm based on the network optimization method in graph theory. X. Liu et al. [19] constructed two distinct multi-objective models: one based on artifact order and the other on batch. These models were developed to address the path optimization problem. It is noticed that the researchers devised an enhanced ant colony optimization algorithm in models. Guo et al. [20] constructed a multi-objective crowd evacuation model based on an improved cuckoo search algorithm, which aims to optimize total time and path length. Zhou X. et al. [21] constructed a TDGVRP model for the time-dependent path planning problem, and an improved ant colony algorithm for calculating the traveling time of vehicles under road allocation strategies. Anshi et al. [22] developed a multi-objective vehicle scheduling model in emergency evacuation based on an enhanced genetic algorithm. Overall, intelligent algorithms were widely used in path optimization and emergency management today.

At present, there is a paucity of research examining the evacuation of rail stations, despite the abundance of studies on large public places, which have primarily focused on timeliness, with fewer of them considering the congestion's effect. However, during the actual evacuation process, evacuees are under a high level of risk-taking as congestion increases. There are two key aspects of the emergency evacuation challenge to be faced: (1) reduce casualties and property losses, which can be achieved by minimizing the total evacuation time, i.e., timeliness; (2) the high-density crowd at the train station tends to panic in an event, which creates congestion and causes the safety risk to blow up. This research applies the adaptive ant colony algorithm into evacuation management considering the congestion degree and obtains a sufficiently effective procedure for evacuation scheme making.

2. Rail Station Personnel Evacuation Model Construction

2.1. Problem Description

Define evacuation paths as all accessible areas for evacuees. In the railway station, each stopover facility is defined as a node, and the channel connected between nodes is defined as a link. The evacuation of a railway station is characterized by dense personnel and limited channels. From the research results of M. Zhou et al. [23], the number of people to be evacuated is N_{ij}, and the ratio of N_{ij} for channel (i,j) and the capacity of channel C_{ij} is defined as the density of personnel $p_{ij}(t)$ at time t, and $p_{ij}(t) = N_{ij}(t)/C_{ij}$, where i denotes the number of the node, j is the channel connection number of the corresponding node, L_{ij} denotes the channel length connected with i and j as end points.

As demonstrated by the findings of He Jianfei and Liu Xiao [24], the evacuation speed is contingent upon the density of individuals involved, and the evacuation process is no longer feasible when density reached a certain value. Furthermore, the obstruction of the people flow can result in overcrowding and trampling incidents. The aforementioned conclusion devotes its attention to the relationship between personnel density and evacuation speed. It defines the relationship between personnel density $p_{ij}(t)$ and evacuation speed $v_{ij}(t)$ on channel (i,j) at time t as follows:

$$v_{ij} = \begin{cases} v_{max} e^{-\alpha p_{ij}(t)} & , p_{ij}(t) > 0.5 \\ v_{max} & , p_{ij}(t) \leqslant 0.5 \end{cases} \quad (1)$$

The effect of crowding on pedestrians is primarily psychological, as evidenced by the observed adjustments in behavior in response to surrounding crowding conditions. Accordingly, when the crowd density is less than 0.5, the speed of pedestrians during evacuation can reach the maximum walking speed. However, when the crowd density is greater than 0.5, the speed of pedestrians during evacuation will be significantly affected. In the case of crowding, a is the correction factor for pedestrian speed during evacuation, where $\alpha = 0.5$.

From Equation (1), if $p_{ij}(t) < 0.5$, it means that the channel (i,j) is smooth, and the channel will not be congested; however, when $p_{ij}(t) > 0.5$ is greater than 0.5, the channel (i,j) will be congested. In order to describe the congestion of each channel more comprehensively in the model, $f_{ij}(t)$ is introduced as the congestion degree of channel (i,j) at time t. In this paper, we define the congestion degree of channel (i,j) at time t as the congestion degree of channel (i,j). In this paper, we define the functional relationship between $p_{ij}(t)$, the density of people at time t, and $f_{ij}(t)$, the congestion degree of channel (i,j), as follows:

$$f_{ij}(t) = \begin{cases} 0 & , p_{ij}(t) < 0.5 \\ e^{\beta p_{ij}(t)} & , p_{ij}(t) \geqslant 0.5 \end{cases} \quad (2)$$

2.2. Multi-Objective Optimization Mathematical Model

The objective of personnel evacuation is to identify an optimal evacuation route within a limited time frame, which can effectively mitigate the probability of adverse events such as congestion and tripping in hazardous area of stations. In conclusion, the primary objective in the planning of the evacuation path is to minimize the evacuation time. Additionally, the degree of congestion during the evacuation process is another optimization objective.

In comparison to other public spaces, railway stations present a unique set of challenges in terms of evacuation procedures. The internal layout is often complex, with a large number of people to be evacuated from a single location. Furthermore, the connection between evacuation routes within and between stations is frequently intricate, and in some instances, there may be no available route or an inoperable route between nodes. It is thus imperative to consider the accessibility of node channels when performing optimal

path optimization. This study presents a multi-objective optimization model of personnel evacuation paths, which is based on the topology of the internal nodes of the station, the attributes of the node channels, and the accessibility between the channels. This model can be effectively applied to solve the objective problem and calculate optimal feasible evacuation paths on a larger scale. The objective of the model is to identify the optimal evacuation strategy for the railway station, whereby the total evacuation time for all individuals is minimized while ensuring a balanced distribution of evacuation loads across the entire system. The pertinent parameter settings of the model are presented in Table 1.

Table 1. Parameters of model.

Notation	Meaning
G	Evacuation network
t	Time
i	Index of node
M	Total number of persons in the danger zone
Pathk	Path of evacuee k
s_0^k	The starting point of the path of evacuee k
l_{ij}	Length of channel (i,j)
u_{ij}	Reachability of channel (i,j), with 0 representing reachable and 1 representing not
$v_{ij}^k(t)$	Velocity of evacuee k in channel (i,j)
$v_{ij}(0)$	Standard velocity of evacuee k in channel (i,j)
$N_{ij}(t)$	Number of persons in channel (i,j) at moment t
C_{ij}	Maximum number of people that channel (i,j) can accommodated
Tmax	Total time for evacuation network to complete full evacuation
α	Correction factor of velocity, taking 0.5
β	Correction factor of congestion, taking 0.5

A mathematical model is introduced as follows:
Objective function:

$$F_1 = \min\left\{\sum_{k=1}^{M} \sum_{i=x_0^k}^{(i,j)\in Pathk} t_{ij}^k\right\},\ F_2 = \min\left\{\sum_{(i,j)\in G} \sum_{t=0}^{T_{max}} f_{ij}\right\} \tag{3}$$

Constraint condition:

$$\begin{aligned}
&v_{ij} = \begin{cases} v_{max} e^{-\alpha p_{ij}(t)} & , p_{ij}(t) > 0.5 \\ v_{max} & , p_{ij}(t) \leqslant 0.5 \end{cases} \\
&f_{ij} = \begin{cases} 0 & , p_{ij}(t) < 0.5 \\ e^{\beta p_{ij}(t)} & , p_{ij}(t) \geqslant 0.5 \end{cases} \\
&\sum_{t=x_0^k}^{(i,j)\in Pathk} u_{ij} = 0 \\
&p_{ij}(t) = \frac{N_{ij}}{C_{ij}} \\
&\int_0^{t_0^k} v_{ij}^k(t)dt = l_{ij} \\
&\frac{N_{ij}}{C_{ij}t} \leq 1
\end{aligned} \tag{4}$$

In this model, we have the following: Equation (3) represents the initial optimization objective, which is to minimize the total evacuation time of all evacuees and represents the secondary optimization objective, which is to minimize the sum of the crowding degree of each channel in the evacuation process. Equation (4) illustrates the correlation between the density of individuals within the channel. Notably, we have further adopted the adaptive Cauchy operator [25] in the experiments. Moreover, it depicts the velocity of individuals traversing the channel in question. It also represents the relationship between density and crowding degree in a given moment. Meanwhile, it imposes constraints on the internal passage of the evacuation path, ensuring its accessibility. Equation (4) also defines the

personnel density, that is, the ratio between the number of people to be evacuated and the number of passages. This provides a link between the passage length l_{ij}, the velocity v(t), and the time t. Furthermore, it constrains the number of evacuees from a passage (i,j) at the same moment to be no higher than the upper limit of the number of people that can be evacuated by the passage.

2.3. Weighted Ideal Point Method

The Weighted Ideal Point Method (WIPM) for multi-objective problems is conducted in the following manner:

$$g(x) = \min\left[\omega_1\left(\frac{f_1 - f_1^{min}}{f_1^{min}}\right)^2 + \cdots + \omega_i\left(\frac{f_i - f_i^{min}}{f_i^{min}}\right)^2 + \cdots + \omega_i\left(\frac{f_i - f_i^{min}}{f_i^{min}}\right)^2\right] \quad (5)$$

where $f_1^{min} = \min_{x\in S} f_i(x)$, i = 1,. . ., k_0 represents the ideal points set, $\omega = (\omega_1,\ldots,\omega_k)$ is weight vector, and $\Sigma_{i=1}^{k}\ \omega_i = 1$, $\omega_i > 0$, i = 1,. . .,k.

In this paper, the WIPM is employed to assign weights to each optimization objective and calculate the comprehensive adaptability according to the weights. This allows the objective function of the mathematical model of people evacuation in railway station to be transformed as follows:

$$F(x) = \min\left[\omega_1\left(\frac{F_1 - F_1^{min}}{F_1^{min}}\right)^2 + \omega_2\left(\frac{F_2 - F_2^{min}}{F_2^{min}}\right)^2\right] \quad (6)$$

In the majority of studies examining evacuation problems, the significance of evacuation time is given greater consideration than congestion. In light of the related studies on a multi-objective evacuation path optimization model by A. Arenas et al. [26] and X. Yang et al. [27], this paper sets *W*1 and *W*2 to 0.7 and 0.3, respectively.

3. Enhanced Quantum Ant Colony Algorithm

3.1. Algorithm Description

The ant colony algorithm is a simulated evolutionary algorithm proposed by the Italian scholars Dorigo and Blum [28]. The algorithm simulates the behavior of ants discovering paths in the process of searching for food, exhibiting characteristics of distributed computation, positive feedback, and heuristic search. It is a widely utilized approach for addressing a range of optimization problems, including the traveler problem, assignment problem, workpiece sequencing problem, and vehicle path problem.

The quantum ant colony algorithm incorporates the benefits of quantum bits into the ant colony algorithm, thereby effectively addressing the limitations of traditional ant colony algorithms, which are known to exhibit slow convergence and a proclivity for local optima. However, due to the extensive scale and dynamic nature of the optimization problem associated with the evacuation of railway stations, the convergence and global requirements of the algorithm are particularly stringent. In light of these considerations, this paper primarily seeks to enhance the quantum ant colony algorithm in three key areas: (1) based on the concept of dynamic self-adaptation, the rotational angle and the amount of rotation in the quantum rotation gates are adaptively enhanced; (2) to guarantee the diversity of the population, the process of ant mutation is incorporated; and (3) ants release pheromones in accordance with the significance of the arc in which they are situated and engage in pheromone updating.

3.2. Enhanced Design of Algorithms

3.2.1. Quantum Rotation Gates

Quantum gates represent a fundamental operational component in the implementation of evolutionary operations. The selection of an appropriate method is often contingent upon

the specific requirements of the problem at hand. In the context of quantum enhancement algorithms, the quantum rotation gate is typically selected as the operator for the evolution process. The following is a definition of a quantum rotation gate:

$$U(\theta_i) = \begin{bmatrix} \cos\theta_i & -\sin\theta_i \\ \sin\theta_i & \cos\theta_i \end{bmatrix} \tag{7}$$

The process of updating a quantum bit is as follows:

$$\begin{bmatrix} \alpha_i' \\ \beta_i' \end{bmatrix} = U(\theta_i) \begin{bmatrix} \alpha_i \\ \beta_i \end{bmatrix} = \begin{bmatrix} \cos\theta_i & -\sin\theta_i \\ \sin\theta_i & \cos\theta_i \end{bmatrix} \begin{bmatrix} \alpha_i \\ \beta_i \end{bmatrix} \tag{8}$$

That is

$$\begin{cases} \alpha_i' = \alpha_i \cos\theta_i - \beta_i \sin\theta_i \\ \beta_i' = \alpha_i \sin\theta_i + \beta_i \cos\theta_i \end{cases} \tag{9}$$

where $|\alpha_i'|^2 + |\beta_i'|^2 = 1$, and the $(\alpha_i,\beta_i)^T$ and $(\alpha_i',\beta_i')^T$ represent the probability of the i-th quantum bit before and after the quantum rotation gate update, respectively. Similarly, θ_i represents the rotation angle, $\Delta\theta_i$ denotes the rotation angle's size, and $s(\alpha_i,\beta_i)$ denotes the rotation angle's direction, $\theta_i = \Delta\theta_i s(\alpha_i,\beta_i)$.

The application of an adaptive strategy to optimize and enhance the ant colony algorithm can markedly enhance the algorithm's performance. The adaptive strategy of the quantum ant colony algorithm is to correlate the value of the rotation angle $\Delta\theta$ with the iteration speed. The value of $\Delta\theta$ exerts an influence not only on the convergence speed of the algorithm but also on the quality of the solution; we have $\Delta\theta = 5\exp(-t/t_{max})$, where t indicates that the algorithm is currently in its t-th iteration. t_{max} means that the algorithm will be executed a maximum of tmax times and then terminated.

The $s(\alpha_i,\beta_i)$ indicates the direction of the rotation angle. This ensures that the algorithm is optimized in the optimal direction. If $s(\alpha_i,\beta_i) > 0$, the rotation gate rotates in a counterclockwise direction; if $s(\alpha_i,\beta_i) < 0$, the rotation gate rotates in a clockwise direction. In this paper, the variable $s(\alpha_i,\beta_i)$ is represented by $s(\alpha_i,\beta_i) = d_i\gamma_i$, where $\gamma_i = \arctan(\beta_{ib}/a_{ib}) - \arctan(\beta_{in}/a_{in})$, and $d_i = a_{in}\beta_{ib}/\beta_{in}\alpha_{ib}\alpha_{ib}$ and β_{1b} represent the probability distributions at the i-th quantum position of the optimal individual. The $\arctan(\beta_{in}/\alpha_{in})$ denotes the current phase at the i-th quantum position of the nth individual, while $\arctan(\beta_{ib}/\alpha_{ib})$ denotes the optimal phase at the -th quantum position of the optimal individual.

3.2.2. Mutations in Individual Ants

In this paper, we build upon the concept of mutation operations in genetic algorithms as proposed by Keshanchi et al. [29] to enhance the diversity of the population in quantum ant colony algorithms. This approach aims to prevent premature convergence of the algorithm. The mutation operation is defined as follows:

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \alpha_{ij} \\ \beta_{ij} \end{pmatrix} = \begin{pmatrix} \alpha_{ij}' \\ \beta_{ij}' \end{pmatrix} \tag{10}$$

where $(\alpha_i,\beta_i)^T$ represents the probability of the j-th bit of the i-th individual before the mutation, whereas $(\alpha_i,\beta_i)^T$ represents the probability of the j-th bit of the i-th individual after the mutation.

3.2.3. Pheromone Updates

The fundamental principle of the ant colony algorithm is the updating of pheromones. In a quantum ant colony algorithm, the quantity of pheromone released is represented by the quantum bit. The matrix B(t) is to be constructed according to Q(t), $B(t) = (b_1^t, b_2^t, \ldots, b_m^t)^T$, where b_i^t is an n-bit binary string of numbers.

$$\Delta\tau_{ab}^{i} = \begin{cases} \frac{Q'}{L_i} & \text{, ant i passesnode b and } p_{ib} = 1 \\ \frac{Q}{L_i} & \text{, ant i passesnode b and } p_{ib} = 0 \\ 0 & \text{, otherwise} \end{cases} \quad (11)$$

If we have $p_{ib} = 1$, this indicates that b represents a critical point; conversely, when $p_{ib} = 0$, it indicates that b does not represent a critical point. Consequently, when $p_{ib} = 1$, the total quantity of pheromone is presumed to be Q', which can be set to $Q' = kQ$. In each iteration, the parameter of the pheromone released on the arc by each ant arriving at each node is recorded by B(t). In the event that the ant traverses the arc and is able to identify the optimal individual on B(t), the pheromone value marking the current arc will be augmented. Conversely, if the ant is unable to locate the optimal individual on B(t), the pheromone value will be diminished. In the absence of ant passage through the current arc, the pheromone value will be set to zero.

3.3. Solving Enhanced Quantum Ant Colony Algorithm

The enhanced quantum ant colony algorithm is employed to address the emergency evacuation path optimization challenge in a railway station as well as the building fire evacuation task. The implementation process encompasses eight principal stages: initialization, ant colony arrangement, binary matrix construction, solution space construction, mutation, evaluation and recording, rotation gate operation, and judgment.

3.4. Toy Example of the Enhanced Quantum Ant Colony Algorithm

In order to ascertain the efficacy of the enhanced quantum ant colony algorithm, Eil51, Eil101, and the Chinese traveler problem dataset from the TSPLIB platform have been selected as the benchmark cases for the algorithm comparison experiments presented in this paper. In parallel, a crowd evacuation network model is developed based on the railway station in W City, which is employed to assess the efficacy of the algorithm in addressing the evacuation challenge. The experiments were programmed and developed in the Matlab 2018 environment, and the code execution and value statistics were conducted on a workstation equipped with an Intel Core i9-12900K CPU and 64 GB RAM. In order to demonstrate the superiority of the enhanced quantum ACO (QACO), the fundamental ACO and chaotic ACO (CACO) are selected for illustrative purposes in this paper as control variables.

At the same time, in order not to lose the generality, each kind of algorithm experiment of each benchmark example is run independently 20 times, and the maximum number of stopping generations of the algorithms are all 200 times. The parameters of the algorithms are set in Table 2, in which the signal generator of the CACO algorithm is $z_{i+1} = 0.4z_i(1 - z_i)$; the quantum multiplier k = 4 for the enhanced QACO algorithm. Notably, the parameters for QACO are further estimated by sensitivity analysis in Section 4.

Table 2. Algorithms parameters.

Algorithms	m	α	β	ρ	Q
ACO	31	1	6	0.9	1
CACO	30	1.5	2	0.9	-
QACO	31	1.5	2	0.9	-

The outcomes of the enhanced QACA are illustrated in Figures 1–4. A comparison of the results yielded by the three algorithms is presented in Tables 3 and 4.

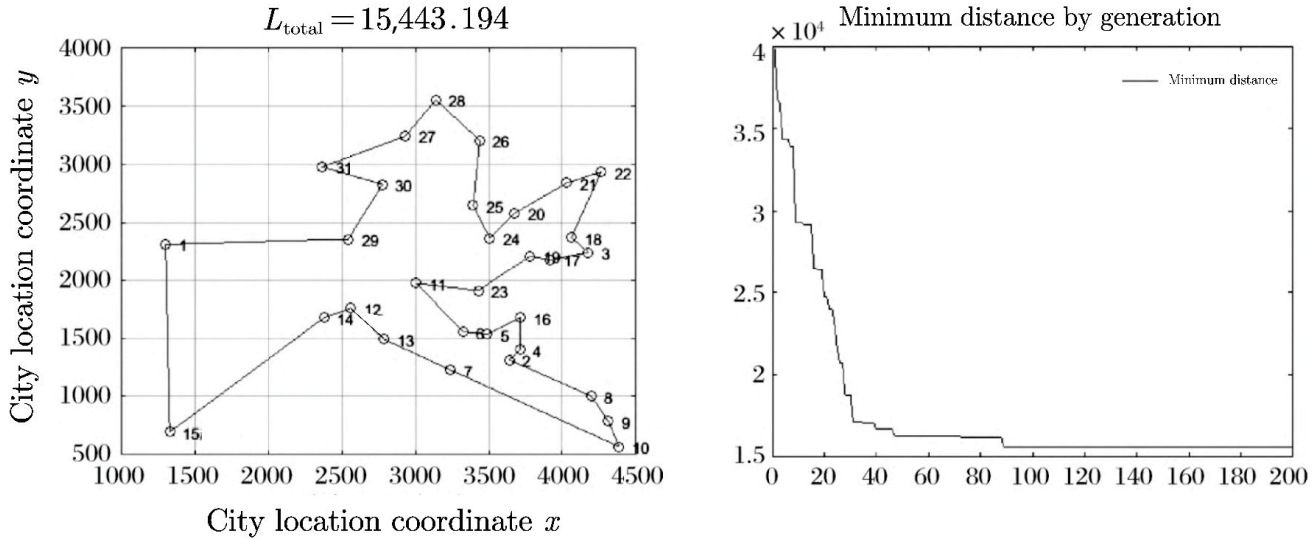


Figure 1. QACA performance for Chinese traveler problem.

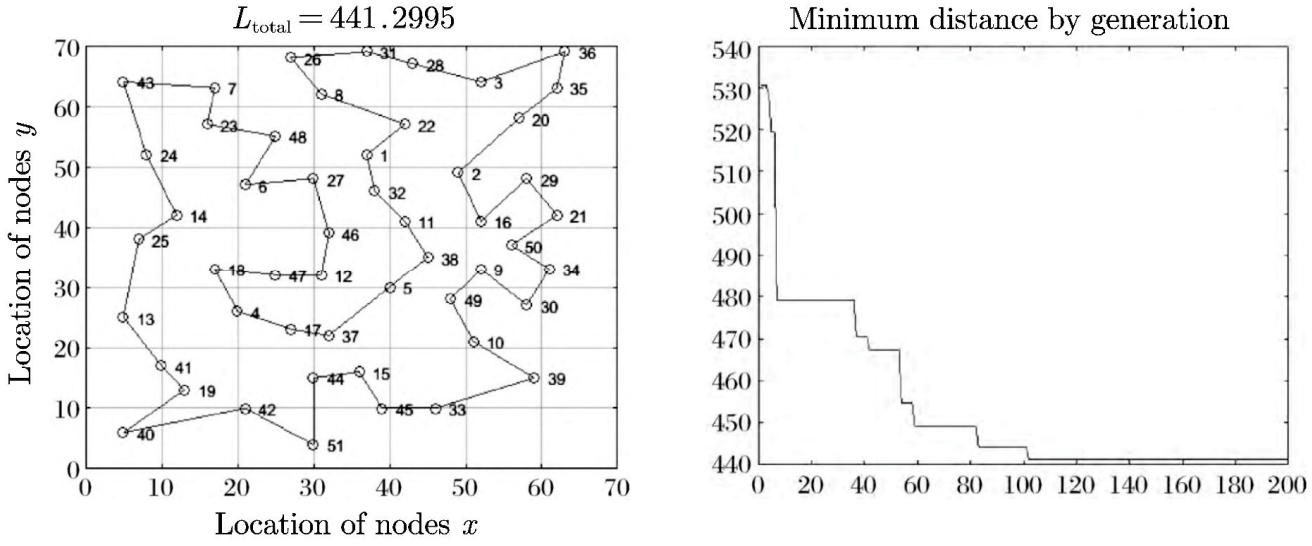


Figure 2. QACA performance for Eil51 problem.

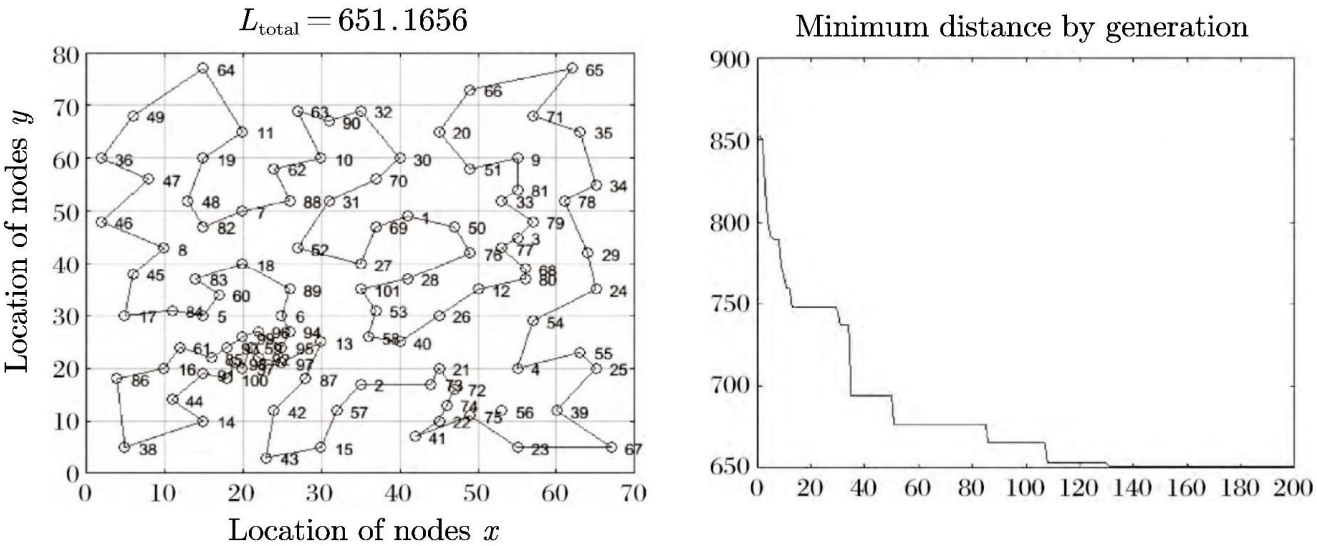


Figure 3. QACA performance for Eil101 problem.

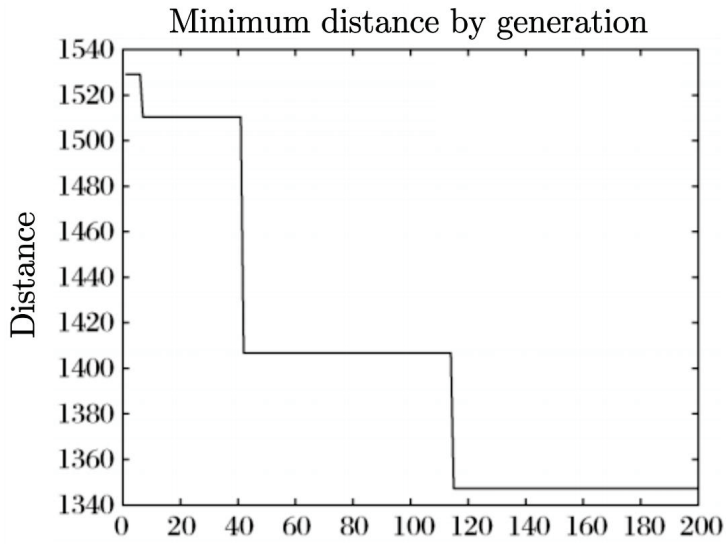


Figure 4. QACA performance for crowd evacuation network model.

Table 3. Comparative analysis of three algorithms on the TSPLIB platform.

Problem	TSPLIB	Algorithm	Best Result	Worst Result	Average
Eil51	426	QCAC	441.30	467.02	458.14
		ACO	457.04	469.41	461.22
		CACO	454.16	467.59	461.87
Eil101	629	QCAC	651.17	702.01	680.59
		ACO	694.97	715.52	700.88
		CACO	665.50	706.94	685.21
CTP	15377	QCAC	15,443.19	15,972.76	15,691.25
		ACO	15,601.92	16,221.34	15,948.52
		CACO	15,592.11	16,247.70	15,813.37

Table 4. Comparative analysis of three algorithms for evacuation in railway station.

Problem	Convergence Iteration Number	Best Result	Worst Result	Average
ACO	154	1392.51	1457.29	1412.12
CACO	175	1375.22	1455.75	1395.78
QACO	117	1347.75	1421.12	1385.44
QACO*	106	1324.18	1406.83	1363.92

A comparison of the results presented in Tables 3 and 4 reveals that QACA exhibits clear advantages in terms of convergence speed and solution quality. This suggests that QACA demonstrates superior comprehensive performance. On the one hand, CACO expands the search space by introducing chaotic variables, which can avoid the algorithm from falling into a local optimum. However, this increases the computation volume of the solution significantly, thus leading to a decrease in algorithm efficiency and solution stability. Nevertheless, the overall performance is still stronger than ACO. Conversely, the aforementioned algorithms demonstrate a negative correlation with problem size in terms of solution quality and convergence speed; that is to say, as problem size increases, solution quality and convergence speed both decline. Conversely, the aforementioned algorithms demonstrate a negative correlation with problem size in terms of solution quality and convergence speed; that is to say, as problem size increases, solution quality deteriorates, and convergence speed slows. Conversely, there is a negative correlation between the aforementioned algorithms and the problem size in terms of solution quality and convergence speed; that is to say, as the problem size increases, the quality of the solution and the speed of convergence both decline.

4. Example Validation of the Model

4.1. Parameter Optimization Strategy

The selection of key parameters in the optimization algorithm is crucial for the stability and effectiveness of the model. To ensure the rationality of the parameter choices, a comprehensive analysis was conducted. To further enhance the performance of the Quantum Ant Colony Optimization (QACO) model, a grid search methodology was employed, inspired by the approach outlined in [30]. This method systematically explored the parameter space to identify the optimal combination of parameters. For completeness, we've conduct experiments with the same m = 32 as stated in Table 5, where Table 6 summarizes the selection range for each parameter.

Table 5. Algorithm initialized parameters with the same m = 31.

Algorithms	m	α	β	ρ	Q
ACO	31	1	6	0.9	1
CACO	31	1.5	2	0.9	-
QACO	31	1.5	2	0.9	-

Table 6. Comparative analysis for evacuation in railway station.

Problem	Convergence Iteration Number	Best Result	Worst Result	Average
ACO	154	1392.51	1457.29	1412.12
CACO	181	1381.30	1456.72	1406.01
QACO	117	1347.75	1421.12	1385.44
QACO*	106	1324.18	1406.83	1363.92

Each parameter combination was tested through multiple iterations to ensure the robustness and reliability of the results. The optimal model, denoted as QACO*, was selected based on its superior performance in terms of both convergence speed and solution quality. In the following experiments, the QACO and QACO* models demonstrated significant improvements in optimizing evacuation time and minimizing congestion compared to the standard QACO model. This enhancement underscores the importance of meticulous parameter tuning in achieving optimal performance for complex optimization problems, such as emergency evacuation in railway stations and fire evacuation among buildings.

Experiments on Railway Station Data

In the event of an emergency at a railway station, the evacuation network comprises 50 nodes. Each node represents a pivotal connection point within the evacuation channel of the station, with the specific configuration varying across different types of stations. The following example illustrates the calculation for a typical high-speed railway station with an evacuee population of M. If the two nodes are able to be interconnected, a straight line is drawn between them, and if both the source node and the target node contain only one element, then the following is true. The pre-evacuation path network diagram is illustrated in Figure 5. In this diagram, $\otimes$ represents the disaster point, the thin solid line represents inter-node access on the second level of the high-speed railway station (HSR), and the dotted line represents access on the first level of the HSR station. The objective of the evacuation task is, therefore, to identify the optimal evacuation path from node 5 to node 29, with the aim of minimizing both the total elapsed time required to complete the evacuation and the cumulative congestion of the path.

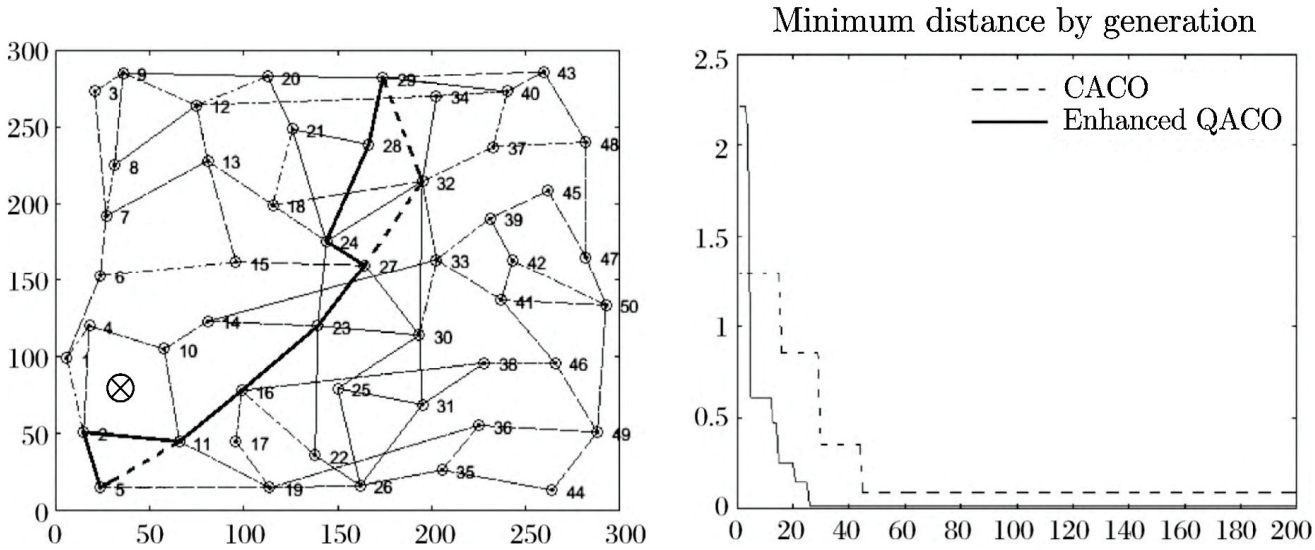


Figure 5. Route result of CACO and QACO.

4.2. Comparative Analysis of Results

4.2.1. Enhanced QACO to CACO

In the case of this HSR station, the enhanced QACO and CACO methods are employed in order to address the issue at hand. For the purposes of this discussion, we will assume that the quantum population size is 31 and that the pheromone importance factor is $\alpha = 1$. The heuristic function importance factor $\beta = 2$, the pheromone volatilization factor $\rho = 0.9$, the quantum multiplicity k = 4, and the number of iterations i = 200. By setting M = 100 and nodes n = 50, the final optimization results, as computed by the enhanced QACO and CACO, are illustrated in Figure 5.

As illustrated in Figure 5, the enhanced QACO algorithm is capable of solving the path A, $5 \rightarrow 11 \rightarrow 16 \rightarrow 23 \rightarrow 27 \rightarrow 32 \rightarrow 29$, whereas the CACO algorithm is better suited to the path B, $5 \rightarrow 2 \rightarrow 11 \rightarrow 16 \rightarrow 23 \rightarrow 27 \rightarrow 24 \rightarrow 28 \rightarrow 29$. A comparison of the two algorithms utilized to solve the aforementioned path is presented in Table 7.

Table 7. Comparative analysis to CACO and QACO.

Problem	Lenth	Total Time	Congestion	Fitness
A	335.942	36,723.63	1.51	0.0936
B	377.227	57,909.89	2.66	0.1866

The results presented in Table 7 and Figure 5 reveals that when the number of iterations is held constant, the enhanced QACO solution path proposed in this study reduces the total length by 41.285 m, the total evacuation time by 21,186.26 s, and the cumulative congestion by 1.15 compared to the CACO algorithm. These findings demonstrate the high solution efficiency and optimization effect of the proposed algorithm, further substantiating its feasibility and superiority.

4.2.2. Comparative Analysis of Evacuation Efficiency with Congestion

In the context of an HSR station, which is a relatively confined space, the number of evacuation routes and the available space are both limited. Consequently, the number of evacuees has a significant impact on the overall evacuation efficiency. To further substantiate the impact of crowding on evacuation efficiency, this study conducted 20 experiments employing the enhanced quantum ant colony algorithm and the hybrid particle swarm algorithm under conditions of varying evacuee numbers (M = 30, 100, 200, and 1000, respectively).

The outcomes of the two algorithms when the number of evacuees is 1000 are illustrated in Figure 5, and the numerical comparison of the fitness function results are presented in Figure 6.

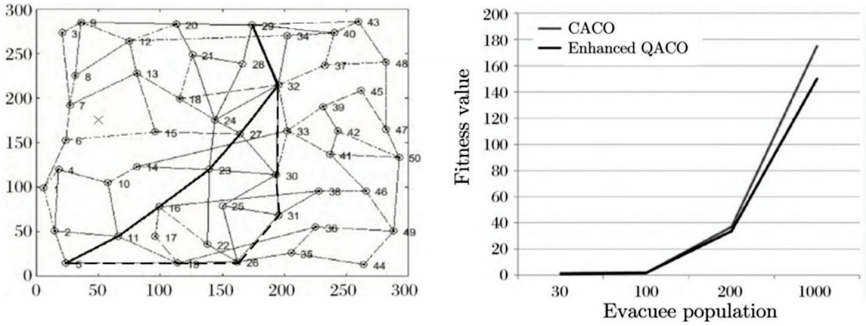


Figure 6. Diagram of CACO and QACO algorithms comparison.

The results of the aforementioned experiments indicate that the enhanced QACO does not exhibit a discernible performance advantage in scenarios where the algorithmic iteration count is limited. Nevertheless, as the number of iterations of the algorithm increases and the complexity of the problem to be solved continues to expand, the enhanced QACO proposed in this paper gradually demonstrates its superiority in terms of convergence speed and solution quality.

4.2.3. Sensitivity Analysis on Evacuation Task

Leveraging the simulation results in Sections 4.2.1 and 4.2.2 and the parameter ranges delineated in Table 8, a systematic sensitivity analysis was conducted as follows.

Table 8. Parameter selection range.

Parameter	Search Range	Description
α	{0.3, 0.5, 0.7}	Correction factor for pedestrian speed, influencing the relation between pedestrian speed and crowd density
β	{0.3, 0.5, 0.7}	Correction factor for congestion degree, influencing the relationship between crowd density and congestion
W_1	{0.50, 0.55, 0.60, 0.65, 0.70, 0.75, 0.80, 0.85, 0.90}	Weight assigned to total evacuation time in the multi-objective optimization
W_2	{0.5, 0.45, 0.40, 0.35, 0.30, 0.25, 0.20, 0.15, 0.10}	Weight assigned to congestion degree in the multi-objective optimization, where $W_2 = 1 - W_1$
Mutation Operator	Simple in Equation (4) or Cauchy Operator	Channel factor on accessibility u_{ij}

Figure 7 presents a three-dimensional response surface that illustrates the coupled sensitivity of α and β with respect to average congestion and the evacuation time. The z-axis represents the normalized composite objective (0 = optimum, 1 = worst). The surface curvature confirms a significant interaction effect (α–β Sobol interaction), validating that simultaneous tuning is indispensable for achieving balanced performance between evacuation time and congestion mitigation.

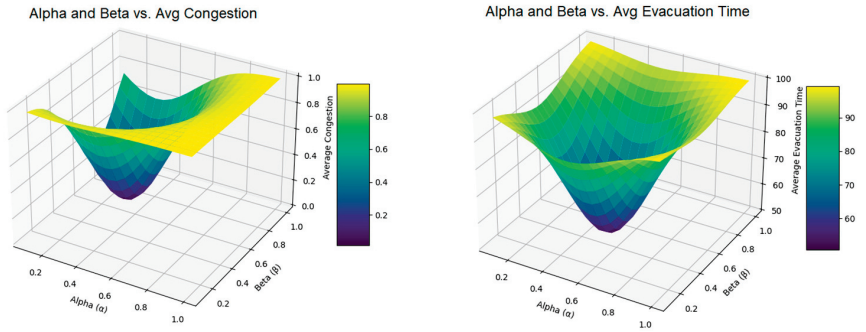


Figure 7. Analysis on the parameters of the enhanced QACO algorithm.

4.3. Experiments on Fire Evacuation Data

To address ecological validity concerns, we implemented a high-fidelity simulation framework adapted from Liang et al.'s building fire evacuation model [30]. For comprehensive sensitivity studies, we conducted full-factorial experiments across critical parameters for our proposed QACO. Based on the article by [31], the data-generation pipeline for the fire-evacuation task can be summarized in two stages: synthetic-scene creation and data pre-processing.

The environment contains a 70 × 70 grid map (1 m^2 cells) with 30% obstacle density (walls/columns) and dynamic hazards on fire propagation, CO toxicity, and crowd density effects. The evacuation data is generated and processed by first tessellating a 5000 m^2 complex into a 70 × 70 occupancy grid, whose black cells represent walls and pillars, then extracting the convex corners of these obstacles as "vertex cells" that constitute the nodes of a reduced graph, whose edges are admitted only when a corridor of at least l m width is unobstructed, compressing the search space.

Results

For comprehensive case studies on both static and dynamic obstacles, we conducted full-factorial experiments across critical parameters for our proposed QACO as follows.

The above results in Tables 9 and 10 verify that dynamic obstacles increased convergence iterations vs. static benchmarks. $W_1 = 0.75$ and $W_2 = 0.3$ are empirically the optimal combination for build QACO*.

Table 9. Experiments on building fire evacuation case with static obstacles.

Problem	Convergence Iteration Number	Best Result	Worst Result	Average
ACO	157	1495.22	1520.64	1508.31
CACO	189	1418.71	1455.75	1431.20
QACO	135	1330.25	1421.12	1398.82
QACO*	122	1329.22	1407.60	1371.30

Table 10. Experiments on building fire evacuation case with dynamic obstacles.

Problem	Convergence Iteration Number	Best Result	Worst Result	Average
ACO	217	1710.34	1745.62	1731.47
CACO	247	1678.26	1703.29	1686.12
QACO	186	1513.49	1617.44	1562.31
QACO*	173	1491.35	1579.31	1538.65

5. Discussion and Conclusions

5.1. Discussions on Future Works

Future research will enhance the model by integrating empirically calibrated panic models (e.g., Helbing–Farkas social-force or HiDAC) to replace the rational-agent assumption and quantify fear-driven route and time deviations, coupling the pedestrian solver to open-source CFD solvers such as the Fire Dynamics Simulator to capture transient fire, smoke, and visibility dynamics for scenario-based optimization of ventilation and signage, and extending the graph representation to 3-D networks that explicitly encode stairs, escalators, elevators, and overlapping floor plates to analyze evacuation in complex multi-level hubs like multi-storey metro stations or airport terminals, collectively yielding a more predictive framework for designers and emergency planners.

5.2. Limitations and Conclusions

This paper examines the emergency evacuation process at a railway station under an emergency situation. From the personnel density, the concepts of congestion degree

and personnel evacuation speed are introduced. Thereafter, the relationship between evacuation efficiency and personnel density, as well as congestion degree, is analyzed. Finally, a multi-objective evacuation path optimization model is established. The model's optimization objectives are twofold: to minimize the total evacuation time and to minimize the cumulative congestion degree of each channel.

Additionally, the evacuation time is considered with the congestion degree, which yields evacuation paths that are more practical and feasible. Moreover, a transformation model is devised to address the shortcomings of a traditional ant colony algorithm, such as being prone to premature convergence, and the new model is well-suited for addressing large-scale path optimization problems.

Then, a comparison between the enhanced quantum ant colony algorithm and the hybrid particle swarm algorithm is made in the context of evacuation crowd simulation. The results demonstrate that with an equivalent number of iterations (e.g., i = 200), the improved quantum ant colony algorithm exhibits a faster convergence, enhances evacuation efficiency by 36.58%, and reduces the congestion of the evacuation process by 43.23%. At the same time, as the number of evacuees increases, so too does the congestion associated with emergency evacuations. This brings a more robust convergence to the enhanced quantum ant colony algorithm. In conclusion, the multi-objective evacuation path optimization model based on the enhanced quantum ant colony algorithm in this paper demonstrates superior efficacy in solving relevant problems.

However, this paper only considers the effective length of the selected channel for the relevant analysis, without considering additional factors such as the distance between the channel and the source of danger, and the panic mentality of the pedestrians. These factors warrant further study. As previously stated, with the ongoing advancement of computer technology, simulation technology is becoming increasingly sophisticated. The question of how to apply software calculations, exemplified by intelligent optimization algorithms, to simulation technology will undoubtedly become a significant area of research.

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

Funding: This research is funded by the National Natural Science Foundation of China (Nos.71774042 and 71271069)and the Heilongjiang Provincial Higher Education Think-Tank (No.ZKKF2022208).

Data Availability Statement: Due to nature of this research, participants of this study did not agree for their data to be shared publicly, so supporting data is not available.

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

References

1. Wu, Y.; Xu, J.; Jia, L.; Qin, Y. Estimation of emergency evacuation capacity for subway stations. *J. Transp. Saf. Secur.* **2018**, *10*, 586–601. [CrossRef]
2. Guo, K.; Zhang, L. Adaptive multi-objective optimization for emergency evacuation at metro stations. *Reliab. Eng. Syst. Saf.* **2022**, *219*, 108210. [CrossRef]
3. Altshuler, E.; Ramos, O.; Núñez, Y.; Fernández, J.; Batista-Leyva, A.J.; Noda, C. Symmetry breaking in escaping ants. *Am. Nat.* **2005**, *166*, 643–649. [CrossRef] [PubMed]
4. Tan, Q.; Huang, G.H.; Wu, C.; Cai, Y. IF-EM: Ans interval-parameter fuzzy linear pro- gramming model for environment-oriented evacuation planning under uncertainty. *J. Adv. Transp.* **2011**, *45*, 286–303. [CrossRef]
5. Cova, T.J.; Johnson, J.P. A network flow model for lane-based evacuation routing. *Transp. Res. Part A Policy Pract.* **2003**, *37*, 579–604. [CrossRef]
6. Yanfeng, L.; Jun, L. The intelligent iterated local search methods based on very large scale neighborhood for vehicle routing problem with fleet size. *Sci. J. Transp.* **2010**, *2*, 47–60.

7. Zhou, M.; Dong, H.; Wen, D.; Yao, X.; Sun, X. Modeling of crowd evacuation with assailants via a fuzzy logic approach. *IEEE Trans. Intell. Transp. Syst.* **2016**, *17*, 2395–2407. [CrossRef]
8. Wang, F.; Pei, Z.; Dong, L.; Ma, J. Emergency resource allocation for multi-period post- disaster using multi-objective cellular genetic algorithm. *IEEE Access* **2020**, *8*, 82255–82265. [CrossRef]
9. Liu, D.; Yu, W.; Baldi, S.; Cao, J.; Huang, W. A switching-based adaptive dynamic programming method to optimal traffic signaling. *IEEE Trans. Syst. Man Cybern. Syst.* **2019**, *50*, 4160–4170. [CrossRef]
10. Li, Y.; Chen, M.; Dou, Z.; Zheng, X.; Cheng, Y.; Mebarki, A. A review of cellular automata models for crowd evacuation. *Phys. A Stat. Mech. Its Appl.* **2019**, *526*, 120752. [CrossRef]
11. Barnes, B.; Dunn, S.; Pearson, C.; Wilkinson, S. Improving human behaviour in macroscale city evacuation agent-based simulation. *Int. J. Disaster Risk Reduct.* **2021**, *60*, 102289. [CrossRef]
12. Guo, X.; Chen, J.; You, S.; Wei, J. Modeling of pedestrian evacuation under fire emergency based on an extended heterogeneous lattice gas model. *Phys. A Stat. Mech. Its Appl.* **2013**, *392*, 1994–2006. [CrossRef]
13. Wang, Y.; Kyriakidis, M.; Dang, V.N. Incorporating human factors in emergency evacuation–An overview of behavioral factors and models. *Int. J. Disaster Risk Reduct.* **2021**, *60*, 102254. [CrossRef]
14. Li, X.; Guo, F.; Kuang, H.; Geng, Z.; Fan, Y. An extended cost potential field cellular automaton model for pedestrian evacuation considering the restriction of visual field. *Phys. A Stat. Mech. Its Appl.* **2019**, *515*, 47–56. [CrossRef]
15. Li, X.J.; Chen, W.B.; Chen, R.X.; Chang, C.T.; Chen, S.H. Evacuation-path-selection model of real-time fire diffusion in urban underground complexes. *Comput. Ind. Eng.* **2023**, *177*, 109014. [CrossRef]
16. Ping, P.; Wang, K.; Kong, D. Analysis of emergency evacuation in an offshore platform using evacuation simulation modeling. *Phys. A Stat. Mech. Its Appl.* **2018**, *505*, 601–612. [CrossRef]
17. Zuo, L.; Shu, L.; Dong, S.; Zhu, C. A multi-objective optimization scheduling method based on the ant colony algorithm in cloud computing. *IEEE Access* **2015**, *3*, 2687–2699. [CrossRef]
18. Zhang, J.; Liu, Y.; Zhao, Y.; Deng, T. Emergency evacuation problem for a multi-source and multidestination transportation network: Mathematical model and case study. *Ann. Oper. Res.* **2020**, *291*, 1153–1181. [CrossRef]
19. Liu, X.; Luo, J. A dynamic multi-objective optimization model with interactivity and uncertainty for real-time reservoir flood control operation. *Appl. Math. Model.* **2019**, *74*, 606–620. [CrossRef]
20. Guo, K.; Zhang, L. Simulation-based passenger evacuation optimization in metro stations considering multi-objectives. *Autom. Constr.* **2022**, *133*, 104010. [CrossRef]
21. Fan, H.; Zhang, Y.; Tian, P.; Lv, Y. Time-dependent multi-depot green vehicle routing problem with time windows considering temporal-spatial distance. *Comput. Oper. Res.* **2021**, *129*, 105211. [CrossRef]
22. Alsnih, R.; Stopher, P.R. Review of procedures associated with devising emergency evacuation plans. *Transp. Res. Rec.* **2004**, *1865*, 89–97. [CrossRef]
23. Zhou, M.; Dong, H.; Zhao, Y.; Ioannou, P. Optimization of crowd evacuation with leaders in urban rail transit stations. *IEEE Trans. Intell. Transp. Syst.* **2019**, *20*, 4476–4487. [CrossRef]
24. Zheng, X.; Cai, L.; Zhang, M.; Jin, L.; Chen, Y. Emergency evacuation path optimization model under multi-export conditions. *China Saf. Sci. J.* **2019**, *29*, 180.
25. Zhang, Y.J.; Wang, Y.F.; Yan, Y.X.; Zhao, J. Self-adaptive hybrid mutation slime mould algorithm: Case studies on uav path planning, engineering problems, photovoltaic models and infinite impulse response. *Alex. Eng. J.* **2024**, *98*, 364–389. [CrossRef]
26. Anaya-Arenas, A.M.; Renaud, J.; Ruiz, A. Relief distribution networks: A systematic review. *Ann. Oper. Res.* **2014**, *223*, 53–79. [CrossRef]
27. Yang, X.; Zhang, R.; Li, Y.; Pan, F. Passenger evacuation path planning in subway station under multiple fires based on multi objective robust optimization. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 21915–21931. [CrossRef]
28. Wang, L.; Niu, Q.; Fei, M. A novel quantum ant colony optimization algorithm and its application to fault diagnosis. *Trans. Inst. Meas. Control* **2008**, *30*, 313–329. [CrossRef]
29. Keshanchi, B.; Souri, A.; Navimipour, N.J. An improved genetic algorithm for task scheduling in the cloud environments using the priority queues: Formal verification, simulation, and statistical testing. *J. Syst. Softw.* **2017**, *124*, 1–21. [CrossRef]
30. Wang, S. Automated Fault Diagnosis Detection of Air Handling Units Using Real Operational Labelled Data Using Transformer-based Methods at 24-hour operation Hospital. *Build. Environ.* **2025**, *282*, 113257. [CrossRef]
31. Liang, J.; Wang, H. Study on building fire evacuation path planning based on improved ant colony algorithm. *J. Syst. Simul.* **2022**, *34*, 1044–1053.

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.

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 978-3-7258-9078-1