water

MDPI

Article

Vision Transformer for Flood Detection Using Satellite Images from Sentinel-1 and Sentinel-2

Ilias Chamatidis [1,*], Denis Istrati [2] and Nikos D. Lagaros [1]

[1] Institute of Structural Analysis and Antiseismic Research, School of Civil Engineering, National Technical University of Athens, GR-15780 Athens, Greece; nlagaros@central.ntua.gr
[2] Department of Water Resources and Environmental Engineering, School of Civil Engineering, National Technical University of Athens, GR-15780 Athens, Greece; distrati@mail.ntua.gr
* Correspondence: ichamatidis@gmail.com

Abstract: Floods are devastating phenomena that occur almost all around the world and are responsible for significant losses, in terms of both human lives and economic damages. When floods occur, one of the challenges that emergency response agencies face is the identification of the flooded area so that access points and safe routes can be determined quickly. This study presents a flood detection methodology that combines transfer learning with vision transformers and satellite images from open datasets. Transformers are powerful models that have been successfully applied in Natural Language Processing (NLP). A variation of this model is the vision transformer (ViT), which can be applied to image classification tasks. The methodology is applied and evaluated for two types of satellite images: Synthetic Aperture Radar (SAR) images from Sentinel-1 and Multispectral Instrument (MSI) images from Sentinel-2. By using a pre-trained vision transformer and transfer learning, the model is fine-tuned on these two datasets to train the models to determine whether the images contain floods. It is found that the proposed methodology achieves an accuracy of 84.84% on the Sentinel-1 dataset and 83.14% on the Sentinel-2 dataset, revealing its insensitivity to the image type and applicability to a wide range of available visual data for flood detection. Moreover, this study shows that the proposed approach outperforms state-of-the-art CNN models by up to 15% on the SAR images and 9% on the MSI images. Overall, it is shown that the combination of transfer learning, vision transformers, and satellite images is a promising tool for flood risk management experts and emergency response agencies.

Keywords: floods; transformer; vision; artificial intelligence; classification

Citation: Chamatidis, I.; Istrati, D.; Lagaros, N.D. Vision Transformer for Flood Detection Using Satellite Images from Sentinel-1 and Sentinel-2. *Water* **2024**, *16*, 1670. https://doi.org/10.3390/w16121670

Academic Editors: Athanasios Loukas, Yaoming Ma and Chang Huang

Received: 29 April 2024
Revised: 2 June 2024
Accepted: 8 June 2024
Published: 12 June 2024

1. Introduction

Every year, floods result in the loss of human lives, the loss of livestock, and millions in economic damages. In recent years, there has been an increase in both the intensity and frequency of these natural disasters, attributed to climate change. The need for forecasting, damage control, and mitigation is of great concern for governments. There are various types of systems and methodologies that can help in forecasting, assessing risk, and predicting damages. This paper will explore the use of artificial intelligence for flood detection through remote sensing images. This can help construct an automated system for flood detection that can assist human action, such as aiding and rescuing affected people in the area. Artificial intelligence can play a crucial role in making these systems faster, cheaper, and more robust. Through its strong predictive capabilities and its ability to discover patterns from large amounts of data, AI can help analyze historical data and make systems that outperform physical and numerical models.

Risk assessment and forecasting systems can help prevent and mitigate damages before they occur and can help in efficiently evacuating people, constructing flood-prevention structures, and reinforcing existing ones. Flood risk assessment systems use historical data to train AI algorithms to create maps indicating the risk of an area. Remote sensing data

and machine learning algorithms, such as SVR, have been used to calculate maps showing the flood inundation depth [1]. Similarly, in [2], several machine learning algorithms, such as random forest and linear regression, are used to assess flood risk, and the information gain ratio is used to calculate the importance of the factors used for training the model (e.g., elevation, curvature, and rainfall). The second type of system used before a flood occurs is the forecasting system. These systems usually monitor streams of data (time series) and try to predict their respective values in a future time window, e.g., the amount of rainfall that will occur in a future 6-h time window. For example, [3] uses LSTM to forecast the flow rate up to a 3-day horizon. This model takes as input a time series of daily discharge and rainfall. Also, in [4], another forecasting system is described that uses a spatiotemporal LSTM, which takes as input hourly rainfall and streamflow data and outputs the outlet flow for the next 6 h. This model also focuses on the interpretability of the results, helping to make more informed decisions.

Apart from the aforementioned techniques used before a flood occurs, there are plenty of systems that are used after a flood. A large portion of these systems focus on flood detection, damage detection, and water body detection. In [5], transformers are used to perform image segmentation in images containing floods, which is a very useful method for automatically detecting water bodies, trees, and houses through aerial imaging. Convolutional neural networks have also been used to assess structural damage in buildings through aerial images from UAVs [6]. Similarly, [7] uses images from UAVs and CNN models to detect damage to infrastructure after a flood. Aerial images and state-of-the-art vision models have been employed to perform scene understanding in images containing flooding events [8]. These models can successfully detect buildings, flooding, and roads, and they can distinguish between floodwater and natural water.

In flood and damage detection cases, models that use AI are very useful as they can make use of images collected from UAVs and satellite images and rapidly analyze complex images. The early detection of floods through remote sensing can be vital in the early deployment of help in that area. Also, robust models that can perform segmentation, mapping, and detection can be used to automatically label huge amounts of data collected daily from various sources. Deep convolutional neural networks have been used to detect floods in Sentinel-2 images [9]. Sentinel-2 contains multiple frequency bands, and the aforementioned study uses green and blue bands, as well as water indices, to make detection easier. Similarly, in [10], images from both Sentinel-1 and Sentinel-2, as well as CNN models, are used to detect floods in these images, achieving an accuracy of 80%. Furthermore, [11] compares several machine learning algorithms (neural networks and SVM) with CNNs for flood detection in radar images. In [12], the performance of a multi-modal model that integrates a CNN with a transformer is compared with that of singular models, such as random forest, neural network, SVM, and CNN. The results show that the transformer combined with the CNN yields better results than the singular models. The authors of [13] compare several segmentation models (WVResU-Net, Swin U-Net, U-Net+++, Attention U-Net, R2U-Net, ResU-Net, TransU-Net, and TransU-Net++) to successfully map flooded areas using Sentinel-1 SAR images. Similarly, in [14], SAR images from Sentinel-1 are used to map inundation extents of lakes. The method uses a CNN to extract high-dimensional features, which are used as input to a transformer, and a fully connected neural network as a classifier head. Moreover, Swin transformers have been used for wetland classification (water, forest wetland, etc.) [15]. The aforementioned study compares the performance of the transformer with CNN models, with the Swin transformer outperforming all other models. Another study [16] uses Sentinel-1 images and Swin transformers to perform water body detection for agricultural reservoirs, while [17] compares Swin transformers with CNNs for wetland classification, using Sentinel-1 and Sentinel-2 images, demonstrating that the former outperforms the latter. Last but not least, [18] combines two models—a Swin transformer and a CNN—to perform water body mapping in remote sensing images. This literature review successfully demonstrates that vision models, especially vision transformers, can be used efficiently for flood detection, segmentation, and mapping.

Also, remote sensing data, such as those from satellites, can be used to train the models because they are rich in information that conventional images cannot capture. This study aims to investigate the possibility of combining transfer learning with vision transformers for fast and automatic detection of flooded areas. This capability is critical for flood risk management agencies and civil protection authorities aiming to quickly decide their emergency response plan after a flooding event. Moreover, although most previous studies focused on applying ViT models on one type of image, the models in this study are applied and evaluated for both SAR and multispectral images from Sentinel-1 and 2, with the objective of developing accurate and "image agnostic" flood detection methodologies that will leverage available data from different sources. Additionally, this study compares the proposed methodology with state-of-the-art CNN models that have shown promise for flood detection applications.

2. Materials and Methods

In this section, the process followed is described. It contains information about the dataset, model, and process used for training.

2.1. Datasets

There are 2 datasets used in this study. Both are part of the SEN12-FLOOD dataset [19], which is an open dataset. It contains 336 time series containing Sentinel-1 and Sentinel-2 images of areas that suffered major flooding events during the winter of 2019. The time span of the acquisition of the images is from December 2018 to May 2019. The observed areas are in East Africa, South-West Africa, the Middle East, and Australia. Each area has multiple images from different times throughout the year, and each image is 512x512 pixels.

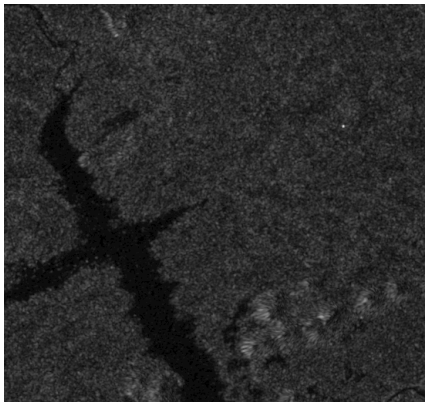
2.1.1. Sentinel-1 Dataset

The Sentinel-1 satellites are a constellation of two polar-orbiting satellites, S1A and S1B, operated by the European Space Agency (ESA). The Sentinel-1 satellites capture radar images of the Earth. Synthetic Aperture Radar (SAR) images have the advantage that they can be acquired in any illumination, even under cloud coverage conditions, which makes it easier to collect data and provide larger amounts of data. Each image in this dataset is in dual-band, containing two polarizations: vertical (VV) and horizontal (VH). The images are provided with radiometric calibration and undergo Range Doppler Terrain Correction using a shuttle-radar topographic mission digital elevation model. For the constitution of the SAR dataset, all available satellite orbits have been considered, resulting in a variety of incidence angles and potential geometric distortions. On average, there are 14 images per sequence. In Table 1, the characteristics of the images captured by Sentinel-1 are analyzed.

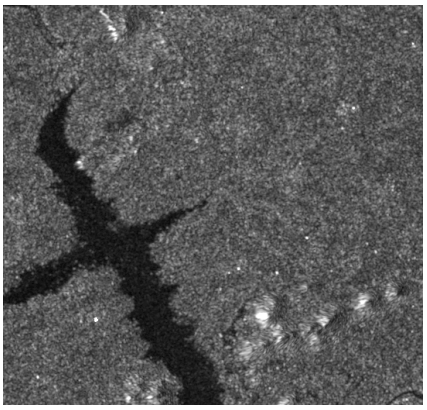
Table 1. Bands captured by Sentinel-2 satellites. Each band is in a different wavelength and has a different pixel size.

Band Name/ Polarization	Pixel Size	Wavelength
VV	10 m	5.405 GHz
VH	10 m	5.405 GHz

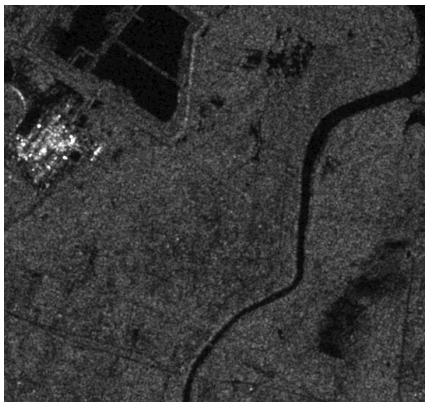
In Figures 1 and 2, there are some examples of images from the Sentinel-1 dataset, with both VH and VV polarizations present. In Figure 1, there are flood events, which are noticeable in the darker areas. In Figure 2, there are no flood events, as the image is more homogeneous with no darker areas.



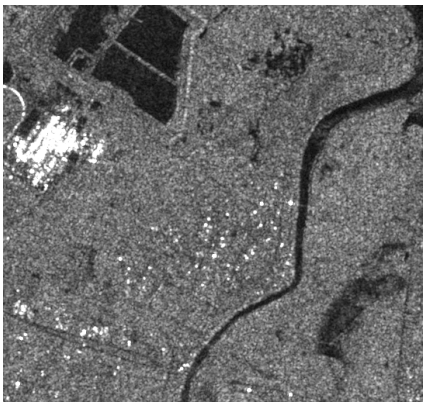
(**a**) Label: Flood; VH polarization



(**b**) Label: Flood; VV polarization

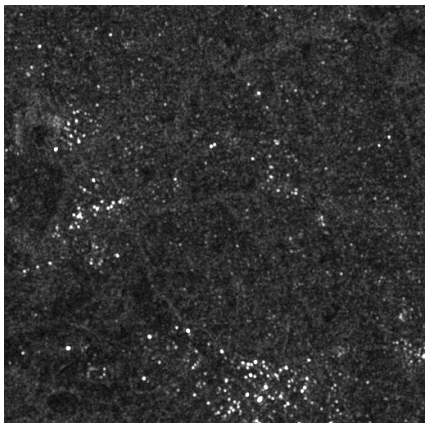


(**c**) Label: Flood; VH polarization

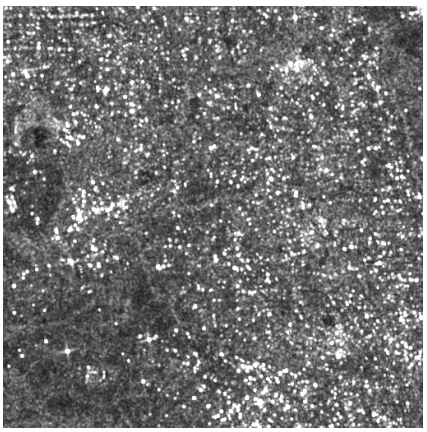


(**d**) Label: Flood; VV polarization

Figure 1. Example of 2 sets of VH and VV polarizations of images containing flood events. Darker areas indicate flooded areas.

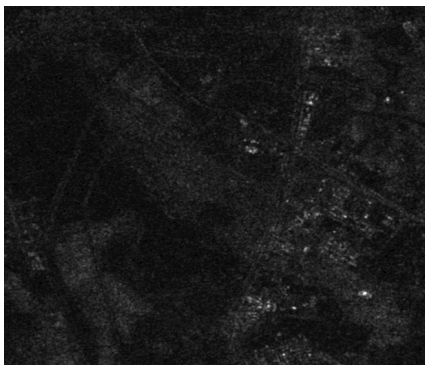


(**a**) Label: No flood; VH polarization

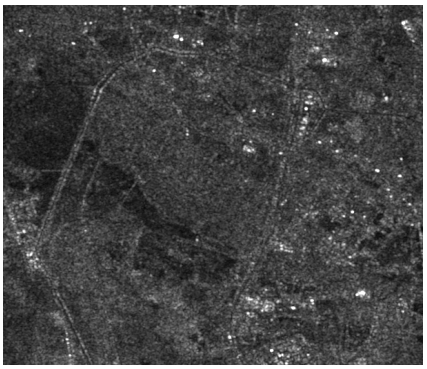


(**b**) Label: No flood; VV polarization

Figure 2. *Cont*.



(**c**) Label: No flood; VH Polarization



(**d**) Label: No flood; VV polarization

Figure 2. Example of 2 sets of VH and VV polarizations of images containing no flood events.

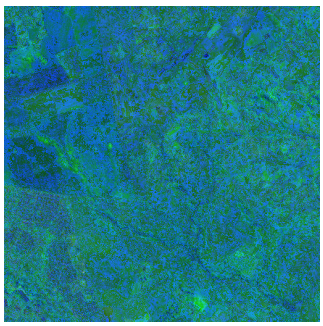
2.1.2. Sentinel-2 Dataset

The Sentinel-2 satellites are a constellation of two polar satellites, S2A and S2B, also operated by the ESA. The Sentinel-2 satellites are equipped with a Multispectral Instrument (MSI) and can provide wide-swath, high-resolution, multispectral images with high temporal sampling, with 9 images per sequence, on average. The Sentinel-2 satellites capture 12 bands. Table 2 presents a summary of the different bands, wavelengths, and pixel sizes that Sentinel-2 satellites can capture.

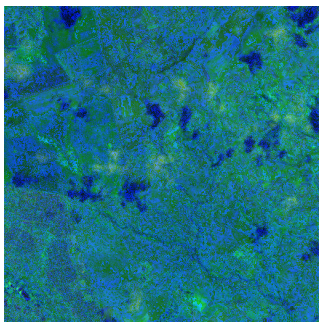
Table 2. Bands captured by Sentinel-2 satellites. Each band is in a different wavelength and has a different pixel size.

Band Name	Pixel Size	Wavelength
B1	60 m	443.9 nm (S2A)/442.3 nm (S2B)
B2	10 m	496.6 nm (S2A)/492.1 nm (S2B)
B3	10 m	560 nm (S2A)/559 nm (S2B)
B4	10 m	664.5 nm (S2A)/665 nm (S2B)
B5	20 m	703.9 nm (S2A)/703.8 nm (S2B)
B6	20 m	740.2 nm (S2A)/739.1 nm (S2B)
B7	20 m	782.5 nm (S2A)/779.7 nm (S2B)
B8	10 m	835.1 nm (S2A)/833 nm (S2B)
B8A	20 m	864.8 nm (S2A)/864 nm (S2B)
B9	60 m	945 nm (S2A)/943.2 nm (S2B)
B11	20 m	1613.7 nm (S2A)/1610.4 nm (S2B)
B12	20 m	2202.4 nm (S2A)/2185.7 nm (S2B)

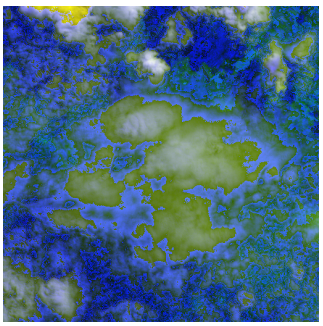
In Figures 3 and 4, there are 2 sets of images showing all the recorded bands. In Figure 3, all the recorded bands for a flooded area are displayed, and in Figure 4, all the bands for a non-flooded area are shown.



(**a**) Label: Flood

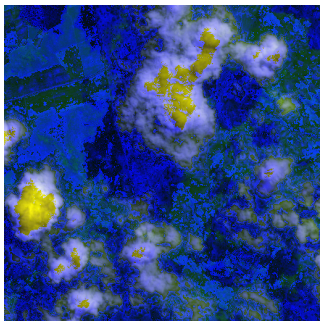


(**b**) Label: Flood

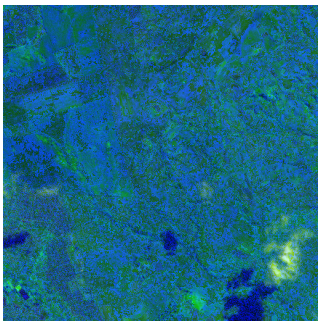


(**c**) Label: Flood

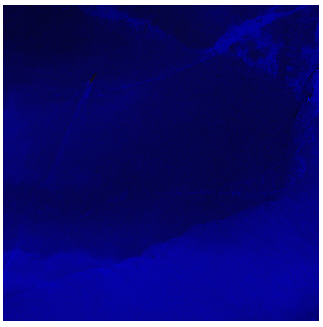
Figure 3. *Cont.*



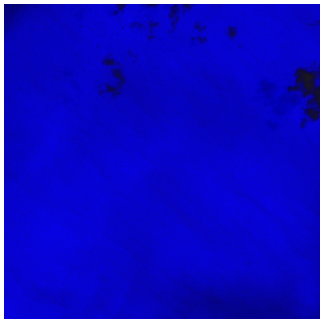
(**d**) Label: Flood



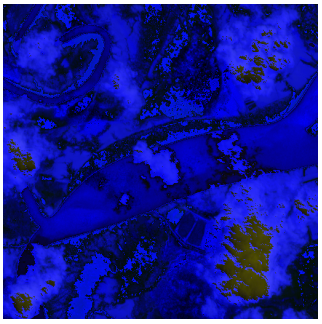
(**e**) Label: Flood



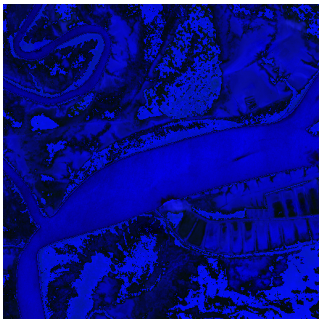
(**f**) Label: Flood



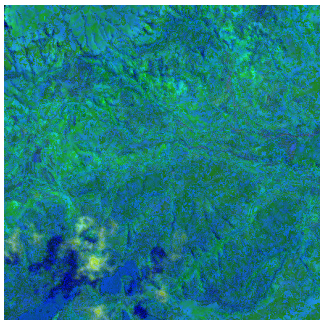
(**g**) Label: Flood



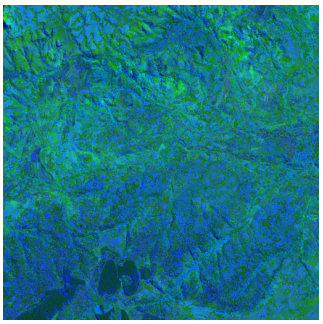
(**h**) Label: Flood



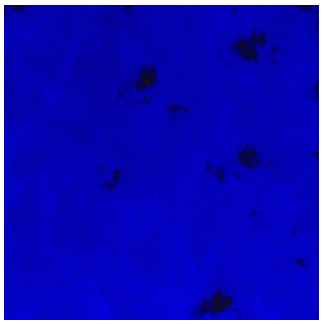
(**i**) Label: Flood



(**j**) Label: Flood

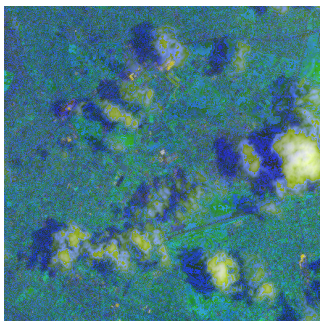


(**k**) Label: Flood

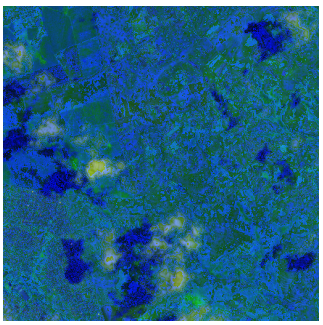


(**l**) Label: Flood

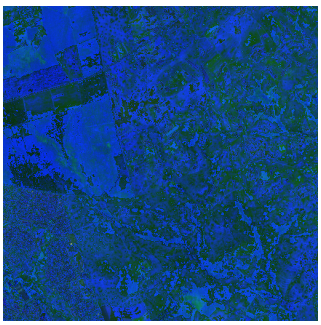
Figure 3. False Sentinel-2 color images of flooded areas.



(**a**) Label: No flood

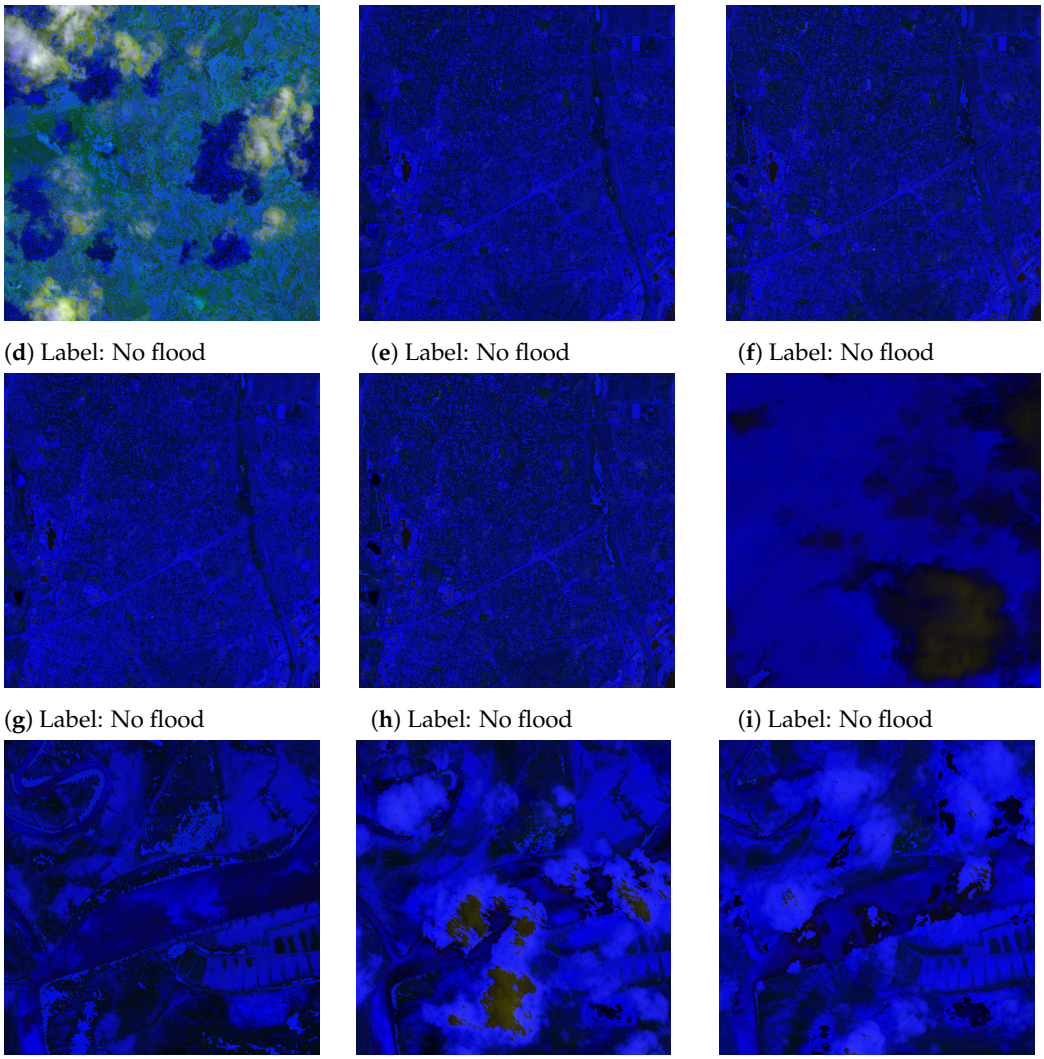


(**b**) Label: No flood



(**c**) Label: No flood

Figure 4. *Cont.*



(**d**) Label: No flood (**e**) Label: No flood (**f**) Label: No flood

(**g**) Label: No flood (**h**) Label: No flood (**i**) Label: No flood

(**j**) Label: No flood (**k**) Label: No flood (**l**) Label: No flood

Figure 4. False Sentinel-2 color images of non-flooded areas.

2.2. Preprocessing

In this section, the methods used to preprocess the datasets are described. Each dataset contains different images; thus, different methods were used.

2.2.1. Sentinel-1

In the Sentinel-1 dataset, there are a total of 3439 images. Each pair of VV and VH images is combined into a 3-channel composite image, with the first channel containing the VV image, the second channel containing the VH image, and the third channel containing the sum of the two, VV + VH. In the metadata of the Sentinel-1 dataset, there are 54 pairs of images that do not exist in the given repository; thus, they are excluded. Furthermore, some faulty images contain mostly black pixels with a value of 0.0. These images are not used in the training and are excluded from the final dataset by using the threshold method, removing images that contain more than 20% values of 0.0. Therefore, 461 images are removed.

The intermediate dataset contains 1105 images with flood events and 1873 images with no flood events. The next step is to balance the dataset so that it contains an equal amount of images from the 2 classes. The chosen number is the count of the smaller class (1105), so using uniform random choice, 1105 images are chosen from the no-flood class. The final preprocessed Sentinel-1 dataset contains 2210 images, balanced equally between the two classes. The next and final step of the preprocessing is to split the dataset into train,

validation, and test sets. The splitting ratio is 0.8, 0.2, and 0.2, respectively. It is important to separate the dataset into 3 sets so that the evaluation of the model uses an independent set (test set) that the model did not "see" during training. The validation set is important for monitoring overfitting and enabling early stopping during training. See Figure 5 for some examples of preprocessed images.

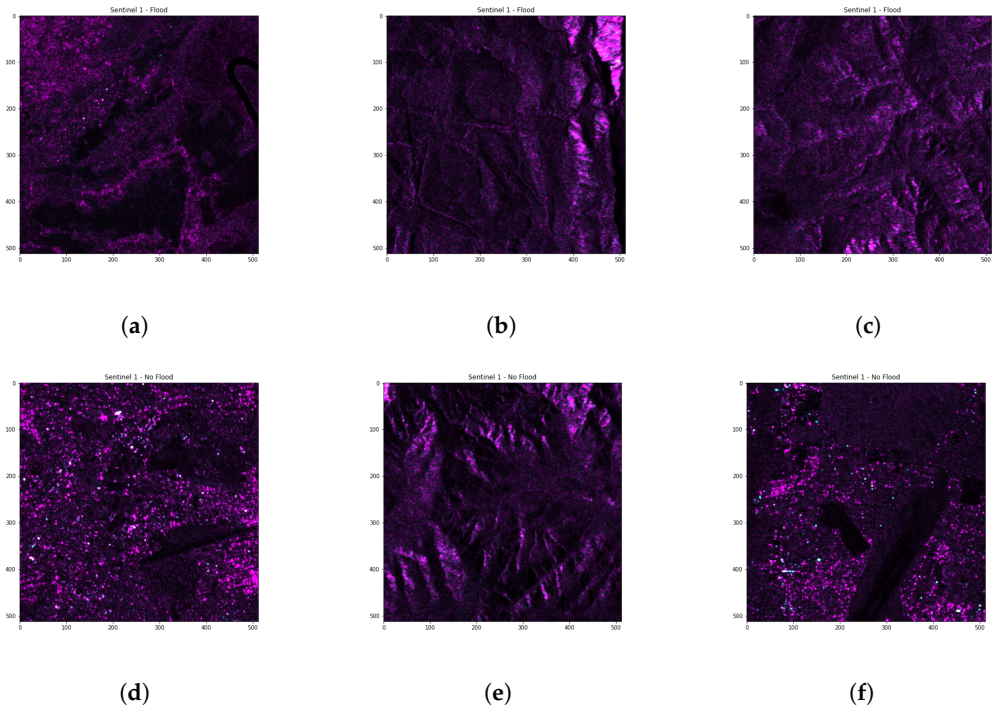


Figure 5. Examples of Sentinel-1 images after preprocessing. Images (**a**–**c**) contain flood events. Images (**d**–**f**) contain non-flood events.

2.2.2. Sentinel-2

In the Sentinel-2 dataset, there are a total of 1973 images. Each image contains the B1, B2, B3, B4, B5, B6, B7, B8, B8A, B9, B11, and B12 bands. A composite image is created with 3 channels: the first channel contains the green band (B08 band), the second channel contains the near-infrared band (NIR, B08 band), and the third channel contains the Normalized Difference Water Index (NDWI). The authors of [20] compare several indices used for water mapping problems and suggest that NDWI yields the best results. It is calculated as shown in Equation (1), where green and nir denote the aforementioned bands and ϵ is a very small value to avoid division by zero:

$$NDWI = (green - nir)/((green + nir) + \epsilon), \quad (1)$$

In Figure 6, there are examples of flood events (upper row) and non-flood events (bottom row). The water is depicted in blue, calculated using the NDWI index. The preprocessed dataset contains 444 images with flood events and 1529 images with non-flood events. To balance the dataset, 444 images with non-flood events are chosen randomly with a uniform distribution, resulting in a final dataset containing 888 images. Also, the dataset is split into 3 subsets: the train, validation, and test sets, with ratios of 0.6, 0.2, and 0.2, respectively.

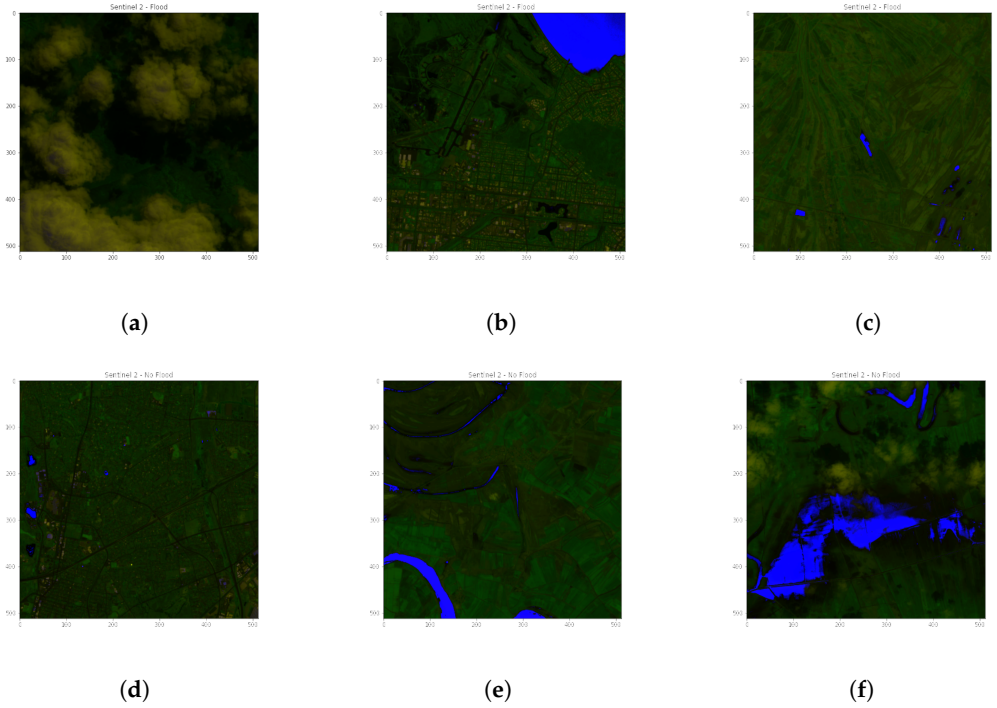


Figure 6. Examples of Sentinel-2 images after preprocessing. Images (**a**–**c**) contain flood events. Images (**d**–**f**) contain non-flood events. Brightness has been increased artificially for visualization purposes.

2.3. Model

Transformers were first introduced in [21] and have since been successfully applied in NLP, machine translation, text generation, and sentiment analysis tasks. They rely on a self-attention mechanism between the tokenized input. They consist of encoder and decoder layers, with the encoder typically used for tasks like text classification, and the decoder for tasks like language generation. Transformers are highly parallelizable, making them suitable for training on large datasets using distributed computing resources.

Vision transformers (ViTs), first introduced in [22], are the adaptations of transformers for image classification tasks. ViTs divide the image into fixed-size patches and treat them as "words" in a sequence. These patches are then passed through a series of transformer layers to capture global and local features. By leveraging self-attention mechanisms, ViTs can effectively model long-range image dependencies, achieving competitive performance on various computer vision tasks. They have shown promising results in tasks such as image classification, object detection, and semantic segmentation, offering a new perspective on visual representation learning. Figure 7 shows the architecture of the model used. The model takes as input a batch of 64 images, which are then converted into patches of size 16×16. These patches are then passed to the linear projection layer and then to the transformer. Finally, they are processed by the MLP in the final layer of the model and are classified into two classes: flood (positive class) and no-flood (negative class).

Transfer learning is a useful method for training when there are insufficient data. Transfer learning involves using pre-trained models that were trained using large amounts of data and then fine-tuning them using your own data, training only the final layer (MLP head). The model used is the pre-trained vision transformer provided by Pytorch [23], which has been pre-trained on thousands of images from ImageNet [24]. Two models are trained, one for each dataset. Each model is trained using the Adam optimizer [25] with learning rates of 0.03, 0.003, 0.0003, and 0.00003. The loss function used is the cross-entropy loss. Both models are trained for 300 epochs, and validation is performed every 10 epochs using the validation dataset. All the layers except the MLP head are frozen because they contain the pre-trained weights; only the MLP head is trained. The MLP head takes as

input the output of the transformer, which has a size of 768 and outputs 2 values (flood and non-flood). The model outputs the logits without passing them through a sigmoid function.

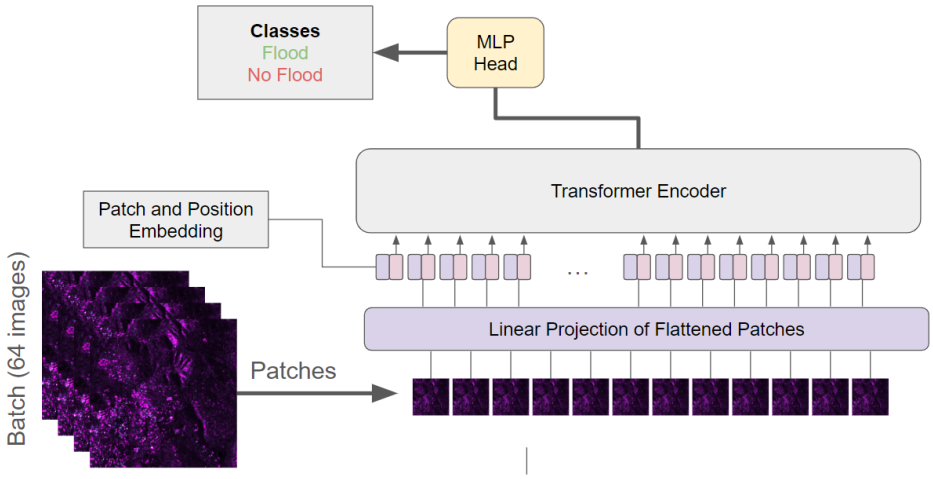


Figure 7. Model architecture of the vision transformer used.

Due to the small size of the dataset, data augmentation is also performed during training. Data augmentation involves artificially increasing the size of the dataset by applying random operations, e.g., rotation and resizing. The operations used to augment the data are random horizontal flip and random vertical flip with $p = 0.5$, which randomly flips an image along its axes. Also, to align with the prerequisites of the pre-trained model, some more operations need to be applied to the image. The first one resizes the image to 256×256, normalizes the image using means of $[0.485, 0.456, 0.406]$ and standard deviations of $[0.229, 0.224, 0.225]$ for each channel, and finally center crops the image to 224×224. Also, the performance of the ViT model is compared with several state-of-the-art CNN vision models. The experimental conditions for these comparisons are the same as those used for the ViT model, including the same dataset, augmentation methods, and number of epochs. Only the last layer of the model is trainable, and the pre-trained model weights [26], trained on Imagenet [24], are used. The models used for comparison are:

- ResNet101 [27];
- VGG16 [28];
- SqueezeNet [29];
- DenseNet121 [30];
- EfficientNet [31];
- MobileNet V2 [32].

3. Results

In this section, the results of the training and testing of the two models are presented, along with some predictions and examples of images.

3.1. ViT Sentinel-1 Results

Figure 8 shows the loss function throughout the training and validation phases. The validation loss indicates that there was no overfitting of the model because it decreased consistently along with the training loss. The model with the best validation loss, observed at epoch 260, was selected for further analysis. Also, the optimal learning rate was determined to be 0.0003.

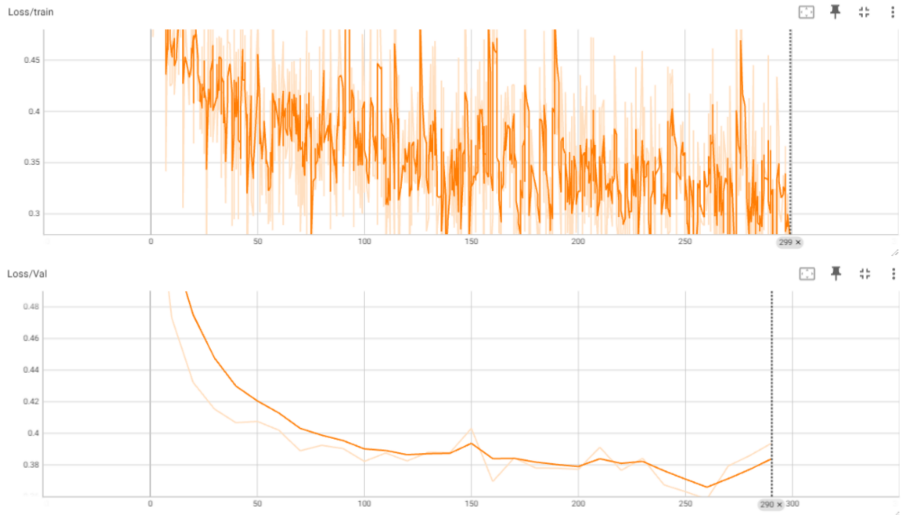


Figure 8. The training and validation losses of the ViT model on the Sentinel-1 dataset for 200 epochs. Validation is performed every 10 epochs.

In Table 3, the results of the ViT model on the Sentinel-1 dataset are presented. The model achieved an accuracy of 84.74%. Table 4 shows the confusion matrix for the model. The rows represent the ground truth, and the columns represent the predicted labels. Notably, the model has 19 cases of False Negatives (FNs), which are flood events classified as non-flood events. In Figure 9, there are four examples of images of flood and non-flood events that were correctly and incorrectly identified in both cases. Also, in Table 5, the performance results of the CNN models are presented.

Table 3. Results for various metrics for the ViT model on the Sentinel-1 dataset.

Metric	Result
Accuracy	84.84%
Recall	81.39%
Precision	91.70%
ROC-AUC	84.58%

Table 4. Confusion matrix for Sentinel-1 ViT model.

	No Flood	**Flood**
No Flood	165	48
Flood	19	210

Table 5. Results of the CNN models on the Sentinel-1 dataset.

Model	Accuracy	Precision	Recall	ROC-AUC
ResNet101	69.68%	67.14%	82.22%	69.24%
VGG16	77.82%	76.30%	82.96%	77.63%
SqueezeNet	80.09%	77.86%	86.02%	79.86%
DenseNet121	77.14%	77.58%	78.60%	77.09%
EfficientNet	79.86%	78.0%	85.15%	79.66%
MobileNet V2	**81.44%**	78.16%	89.08%	81.16%

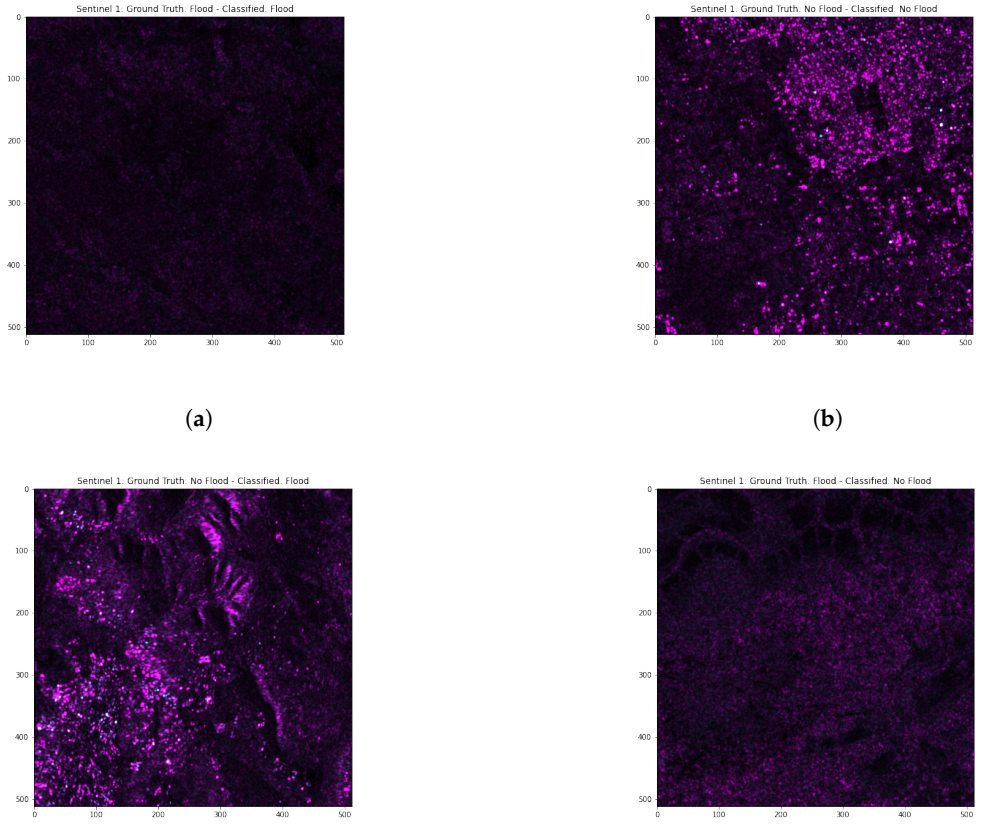


Figure 9. Examples of images that were correctly and incorrectly identified. (**a**) Correctly identified image containing a flood event (TP). (**b**) Correctly identified image containing a non-flood event (TN). (**c**) Non-flood image identified as a flood event (FP). (**d**) Flood event identified as a non-flood event (FN).

3.2. ViT Sentinel-2 Results

In this section, the results of the training and testing of the two models are presented, along with some predictions and examples of images.

3.3. ViT Sentinel-1 Results

Figure 10 shows the loss function throughout the training and validation phases. The validation loss indicates that there was no overfitting of the model because it decreased consistently along with the training loss. The model with the best validation loss, observed at epoch 220, was selected for further analysis. Also, the optimal learning rate was determined to be 0.0003.

Table 6 shows the results of the ViT model on the Sentinel-1 dataset. The model achieved an accuracy of 84.74%. Table 7 shows the confusion matrix for the model. The rows represent the ground truth, and the columns represent the predicted labels. Notably, the model has 12 cases of False Positives (FPs), which are flood events classified as non-flood events. In Figure 11, there are four examples of images of flood and non-flood events that were correctly and incorrectly identified in both cases.

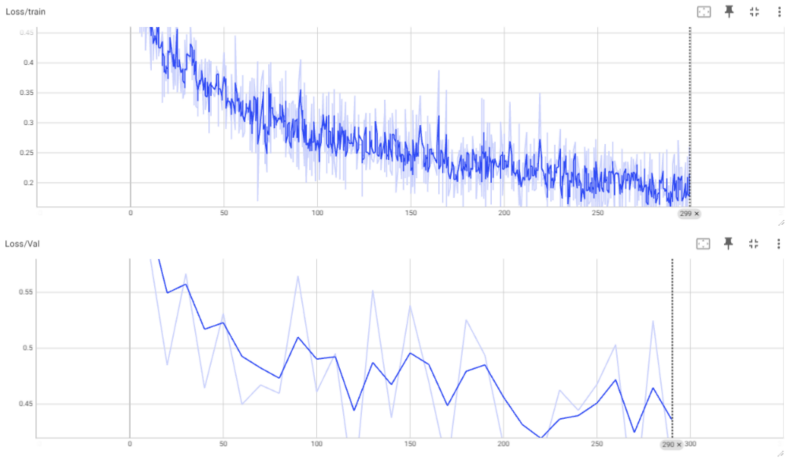
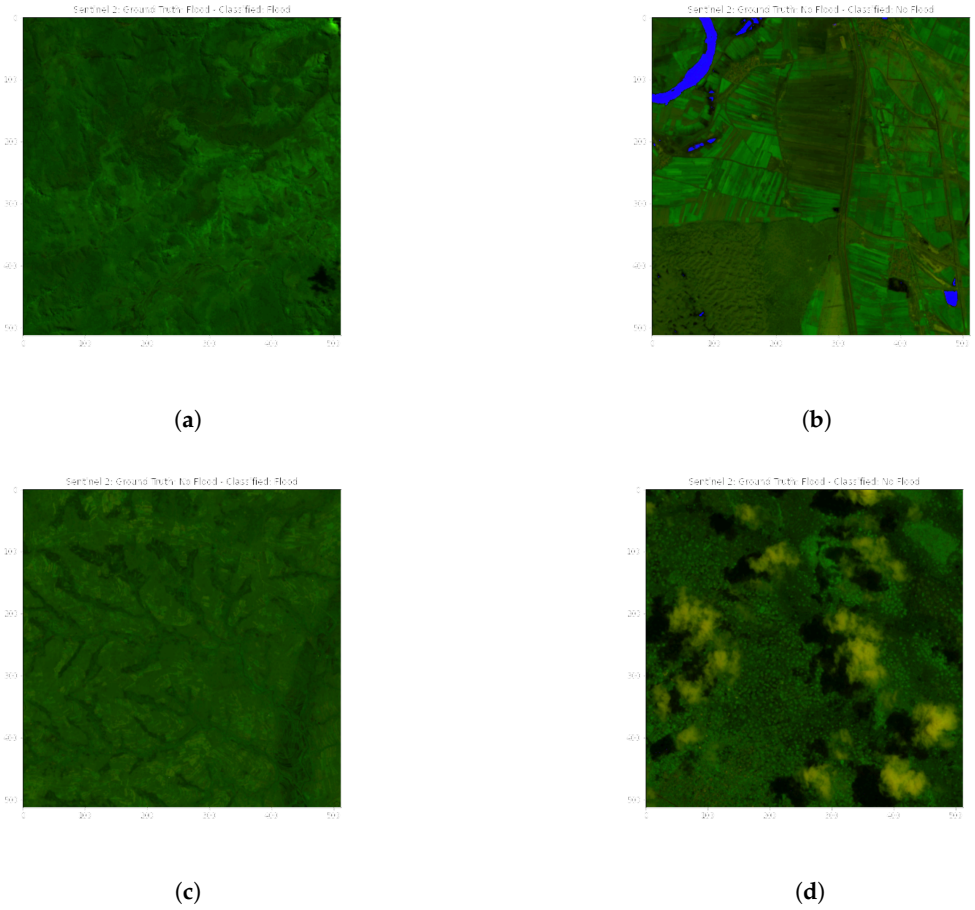


Figure 10. The training and validation losses of the ViT model on the Sentinel-2 dataset for 200 epochs. Validation is performed every 10 epochs.



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

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

Figure 11. Examples of images that were correctly and incorrectly identified (brightness of the image has been increased for visualization purposes). (**a**) Correctly identified image containing a flood event (TP). (**b**) Correctly identified image containing a non-flood event (TN). (**c**) Non-flood image identified as a flood event (FP). (**d**) Flood event identified as a non-flood event (FN).

Also, in Table 8, the performance results of the CNN models are presented.

Table 6. Results for various metrics for the ViT model on the Sentinel-2 dataset.

Metric	Result
Accuracy	83.14%
Recall	84.81%
Precision	78.82%
ROC-AUC	82.96%

Table 7. Confusion matrix for Sentinel-2 ViT model.

	No Flood	Flood
No Flood	81	12
Flood	18	67

Table 8. Results of the CNN models on the Sentinel-2 dataset.

Model	Accuracy	Precision	Recall	ROC-AUC
ResNet101	81.46%	80.95%	80.0%	81.39%
VGG16	79.77%	79.51%	77.64%	79.68%
SqueezeNet	75.28%	74.69%	72.94%	75.18%
DenseNet121	81.46%	79.54%	82.35%	81.49%
EfficientNet	**84.83**%	83.72%	84.70%	84.82%
MobileNet V2	79.21%	80.0%	75.29%	79.04%

4. Discussion and Conclusions

In this study, an approach using a vision transformer and transfer learning is used to detect flooding in satellite images. Two datasets are used to train two different models. The first dataset is the Sentinel-1 dataset, which contains SAR images from flood and non-flood events from different areas. The second dataset is the Sentinel 2 dataset, which contains multispectral images from several flood and non-flood events. The model used is a vision transformer, a sub-category of transformers that has been successfully applied in NLP tasks and shows promise for vision classification tasks due to the encodings and high-level representations that can be leveraged for complex image tasks. Both models achieved similar performance, with high accuracies of 84.84% and 83.14% on the Sentinel-1 and Sentinel-2 datasets, respectively. Also, the ViT model outperformed all the CNN models on the Sentinel-1 dataset by as much as 15%. On the Sentinel-2 dataset, the ViT model outperformed all the models by as much as 9%, but it did not outperform EfficientNet, which yielded similar results (less than a 2% difference). The present study reveals the ability of the vision transformer to automatically detect flooding from different types of satellite data (SAR, multispectral). Another ability of this model is the minimal preprocessing done to the images, which demonstrates the strong ability of the model to detect connections between the input and output. Also, the advantages of transfer learning are successfully demonstrated, which is useful when there are insufficient data available for training the model. Finally, these systems are fast when used for inferring data and can be used for real-life applications with minimal processing power. Overall, this study proves that digital technologies, such as a combination of remote sensing and vision transformers, can be indispensable tools for fast flood mapping by civil protection agencies and emergency response execution. Future studies should further utilize vision transformer models for flood detection, scene segmentation, and understanding damage detection on structures and the environment. Another limitation of this study was that the satellite images were temporally sparse, which restricted the use of the proposed methodology solely to flood detection. However, it is expected that if temporally continuous images (e.g., every 3 h, 6 h, 12 h) become available, the presented methodology could be extended to flood forecasting as well.

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

Funding: This research was supported by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the "2nd Call for H.F.R.I. Research Projects to support Faculty Members and Researchers", AMOSS project: "Additively Manufactured Optimized 3D Printed Steel Structures", (Project Number: 02779).

Data Availability Statement: The data [19] used in the present study are openly available and free to use. The data are available at https://ieee-dataport.org/open-access/sen12-flood-sar-and-multispectral-dataset-flood-detection (accessed on 6 June 2024).

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

Abbreviations

The following abbreviations are used in this manuscript:

AUC	Area Under Curve
FN	False Negative
FP	False Positive
MSI	Multispectral Instruments
MLP	Multi-Layer Perceptron
NDWI	Normalized Difference Water Index
NIR	Near Infrared
NLP	Natural Language Processing
ROC	Receiver Operating Characteristic
TN	True Negative
TP	True Positive
ViT	Vision Transformer

References

1. Mojaddadi, H.; Pradhan, B.; Nampak, H.; Ahmad, N.; Ghazali, A.H.b. Ensemble machine-learning-based geospatial approach for flood risk assessment using multi-sensor remote-sensing data and GIS. *Geomat. Nat. Hazards Risk* **2017**, *8*, 1080–1102. [CrossRef]
2. Chapi, K.; Singh, V.P.; Shirzadi, A.; Shahabi, H.; Bui, D.T.; Pham, B.T.; Khosravi, K. A novel hybrid artificial intelligence approach for flood susceptibility assessment. *Environ. Model. Softw.* **2017**, *95*, 229–245. [CrossRef]
3. Le, X.H.; Ho, H.V.; Lee, G.; Jung, S. Application of long short-term memory (LSTM) neural network for flood forecasting. *Water* **2019**, *11*, 1387. [CrossRef]
4. Ding, Y.; Zhu, Y.; Feng, J.; Zhang, P.; Cheng, Z. Interpretable spatio-temporal attention LSTM model for flood forecasting. *Neurocomputing* **2020**, *403*, 348–359. [CrossRef]
5. Roy, R.; Kulkarni, S.S.; Soni, V.; Chittora, A. Transformer-based Flood Scene Segmentation for Developing Countries. *arXiv* **2022**, arXiv:2210.04218.
6. Gulgec, N.S.; Takáč, M.; Pakzad, S.N. Structural damage detection using convolutional neural networks. In *the Model Validation and Uncertainty Quantification, Volume 3: Proceedings of the 35th IMAC, A Conference and Exposition on Structural Dynamics 2017*; Springer: Cham, Switzerland, 2017; pp. 331–337.
7. Munawar, H.S.; Ullah, F.; Qayyum, S.; Khan, S.I.; Mojtahedi, M. UAVs in disaster management: Application of integrated aerial imagery and convolutional neural network for flood detection. *Sustainability* **2021**, *13*, 7547. [CrossRef]
8. 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]
9. Jain, P.; Schoen-Phelan, B.; Ross, R. Automatic flood detection in SentineI-2 images using deep convolutional neural networks. In Proceedings of the 35th Annual ACM Symposium on Applied Computing, Brno, Czech Republic, 30 March–3 April 2020; pp. 617–623.
10. Bhadra, T.; Chouhan, A.; Chutia, D.; Bhowmick, A.; Raju, P. Flood detection using multispectral images and SAR data. In *Proceedings of the International Conference on Machine Learning, Image Processing, Network Security and Data Sciences, Silchar, India, 30–31 July 2020*; Springer: Singapore, 2020; pp. 294–303.
11. Islam, K.A.; Uddin, M.S.; Kwan, C.; Li, J. Flood detection using multi-modal and multi-temporal images: A comparative study. *Remote Sens.* **2020**, *12*, 2455. [CrossRef]

12. Jamali, A.; Mahdianpari, M. Swin transformer and deep convolutional neural networks for coastal wetland classification using sentinel-1, sentinel-2, and LiDAR data. *Remote Sens.* **2022**, *14*, 359. [CrossRef]
13. Jamali, A.; Roy, S.K.; Beni, L.H.; Pradhan, B.; Li, J.; Ghamisi, P. Residual wave vision U-Net for flood mapping using dual polarization Sentinel-1 SAR imagery. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *127*, 103662. [CrossRef]
14. Dong, Z.; Liang, Z.; Wang, G.; Amankwah, S.O.Y.; Feng, D.; Wei, X.; Duan, Z. Mapping inundation extents in Poyang Lake area using Sentinel-1 data and transformer-based change detection method. *J. Hydrol.* **2023**, *620*, 129455. [CrossRef]
15. Jamali, A.; Mahdianpari, M. Swin transformer for complex coastal wetland classification using the integration of Sentinel-1 and Sentinel-2 imagery. *Water* **2022**, *14*, 178. [CrossRef]
16. Choi, S.; Youn, Y.; Kang, J.; Kim, S.; Jeong, Y.; Im, Y.; Seo, Y.; Kim, W.; Choi, M.; Lee, Y. Waterbody detection for the reservoirs in South Korea using Swin Transformer and Sentinel-1 images. *Korean J. Remote Sens.* **2023**, *39*, 949–965.
17. Jamali, A.; Mohammadimanesh, F.; Mahdianpari, M. Wetland classification with Swin Transformer using Sentinel-1 and Sentinel-2 data. In Proceedings of the IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium, Kuala Lumpur, Malaysia, 17–22 July 2022; pp. 6213–6216.
18. Li, J.; Li, G.; Xie, T.; Wu, Z. MST-UNet: A modified Swin Transformer for water bodies' mapping using Sentinel-2 images. *J. Appl. Remote Sens.* **2023**, *17*, 026507. [CrossRef]
19. Rambour, C.; Audebert, N.; Koeniguer, E.; Le Saux, B.; Crucianu, M.; Datcu, M. SEN12-FLOOD: A SAR and Multispectral Dataset for Flood Detection. 2020. Available online: https://ieee-dataport.org/open-access/sen12-flood-sar-and-multispectral-dataset-flood-detection (accessed on 6 June 2024).
20. Zhou, Y.; Dong, J.; Xiao, X.; Xiao, T.; Yang, Z.; Zhao, G.; Zou, Z.; Qin, Y. Open surface water mapping algorithms: A comparison of water-related spectral indices and sensors. *Water* **2017**, *9*, 256. [CrossRef]
21. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017.
22. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
23. 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*; Curran Associates, Inc.: New York, NY, USA, 2019; pp. 8024–8035.
24. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255. [CrossRef]
25. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
26. Albardi, F.; Kabir, H.D.; Bhuiyan, M.M.I.; Kebria, P.M.; Khosravi, A.; Nahavandi, S. A comprehensive study on torchvision pre-trained models for fine-grained inter-species classification. In Proceedings of the 2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Melbourne, Australia, 17–20 October 2021; pp. 2767–2774.
27. 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, 26 June–1 July 2016; pp. 770–778.
28. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
29. Iandola, F.N.; Han, S.; Moskewicz, M.W.; Ashraf, K.; Dally, W.J.; Keutzer, K. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size. *arXiv* **2016**, arXiv:1602.07360.
30. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
31. Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In Proceedings of the International Conference on Machine Learning PMLR, Long Beach, CA, USA, 10–15 June 2019; pp. 6105–6114.
32. 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 City, UT, USA, 18–23 June 2018; pp. 4510–4520.