A Deep Learning Ensemble Model for Flood Image Classification

Asghar Ali Chandio [1,*], Mehwish Leghari [2], Sahil Umar [1]

[1] Artificial Intelligence Department, Quaid-e-Awam University of Engineering Science & Technology, Nawabshah
[2] Data Science Department, Quaid-e-Awam University of Engineering Science & Technology, Nawabshah
* Correspondence: asghar.ali@quest.edu.pk

Abstract

Flood is a type of natural disaster that leads to a widespread devastation. The increasing amount of rain specifically in the Urban regions of Sindh province causes several issues, whereas the drainage system is not very efficient to handle the large amount of water in a short period of time. Identification of floods is essential for disaster response, as it helps locate areas which need immediate help. Recently, the deep learning-based models have shown the best performance for image classification tasks. In this paper, a deep learning-based ensemble model has been developed where four state-of-the-art deep learning models are combined to classify flood from the images either captured with the mobile camera or other image capturing devices. The deep learning ensemble model has been trained and tested on the two publicly available datasets labelled with flood and non-flood images. To enhance the efficacy of the deep learning-based ensemble model, the hyper-parameters of the four models are fine-tuned. The results obtained show that the deep learning-based ensemble model outperforms than the individual models.

Keywords: Flood Classification; Deep Ensemble Model; Urban Flood Classification

1. Introduction

Floods are among the most devastating natural disasters, often leading to significant loss of life, destruction of property, and disruption of livelihoods [1, 2]. In recent years, the frequency and intensity of floods have escalated, particularly in urban regions of Sindh province, Pakistan. These areas face com- pounded challenges due to inadequate drainage systems that struggle to manage the large volumes of water resulting from heavy rainfall within short periods. The timely identification and classification of floods are crucial for effective disaster response and resource allocation, enabling rapid assistance to affected areas. With advancements in technology, image-based flood identification has emerged as a powerful tool for disaster management. Leveraging the capabilities of deep learning, a field that has revolutionized image classification tasks, this research work explores the development of a robust solution for flood classification from flood and non-flood images. Deep learning models with ensemble learning have demonstrated exceptional results in processing and analyzing complex visual data, making them suitable to accurately differentiate between flood scenes from non-flood images captured with the help of various smart devices, including digital cameras of a mobile phone. This research work uses a deep learning-based ensemble method designed to enhance the accuracy and reliability of flood classification from images. The ensemble models have not been much used for flood image classification, while these models have shown state-of-the-art accuracies in other image classification tasks [3–5]. Furthermore, the deep feature fusion and data augmentation methods have also been commonly used in image classification and recognition problems [6, 7]. In this research work, the deep ensemble model integrates

four different deep learning architectures, each contributing unique strengths to the efficiency of the overall model. By fusing these models, the ensemble model leverages their complementary features, which results in higher performance as compared to the individual models. To further optimize the efficacy of the ensemble model, the tunning of the hyper-parameters has meticulously been performed for each constituent model. Due to the unavailability of the flood and non-flood image datasets of Sindh province or Pakistan, two different publicly available datasets including Flood IMG: Flood Image Database System [8] and the CHF2015 dataset Images [9] were used to train and evaluate the performance of the individual deep learning models and the deep ensemble models. The datasets contain both flood and non-flood images. Some non-flood images from Google images were also downloaded and used to train the deep learning models. The results of the experiments illustrate that the deep ensemble models can perform better than the individual models. However, the deep ensemble models require more data for training and model generalization.

2. Proposed Methodology

Flood and non-flood classification from images using deep learning-based ensemble models was performed through a structured approach. The methodology began by collecting the flood and non-flood images from two publicly available datasets, which consisted of diverse images of flood and non-flood regions, and ensured representation of data with different variations. After reading data, all the images were resized to a fixed size of 224x224x3. Some comprehensive preprocessing techniques including resizing, data normalization, and data augmentation were performed in order to enhance model generalization capability [10]. The important features distinguishing the flood and non-flood areas were extracted using the convolutional neural network. Furthermore, the main features extracted by each individual deep learning model were given to the deep ensemble model for its state-of-the-art performance in image classification tasks such as flood and non-flood. The four models of deep learning were fine-tuned by changing their hyper-parameters to optimise and maximise the contribution of individual model to the overall performance of ensemble models. Finally, the ensemble model was trained and tested using rigorous metrics to validate its performance over the individual deep learning models. Figure 1 illustrates a flowchart representing the methodology followed in this research.

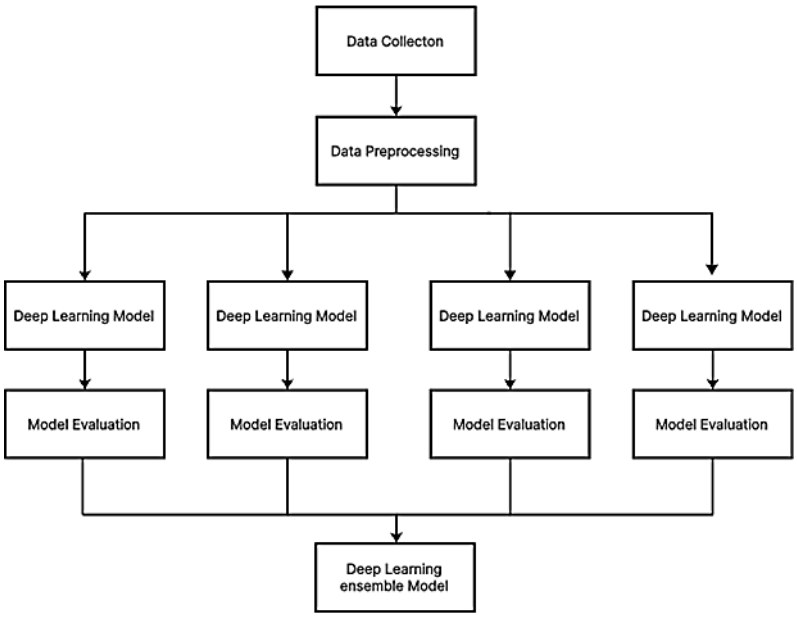

Figure 1. Proposed Methodology of Flood and Non-Flood Image Classification using Deep Learning Ensemble Model

2.1. Data Collection

In order to train and evaluate the performance of individual and deep learning- based ensemble models, two datasets were collected from Kaggle [8, 9], containing annotated images of flood and non-flood scenes. The datasets are publicly available. The datasets contain diverse and representative data, forming a robust foundation for training and testing the deep learning ensemble model.

Dataset Limitation: This study relied solely on publicly available datasets (FloodIMG and CHF2015), and some non-flooded images were downloaded from Google, which may not fully capture the unique visual features of floods in the Sindh region or other parts of Pakistan. The lack of region-specific imagery may limit the generalizability of the model to local disaster scenarios. To address this, future research should consider building a large-scale, locally curated dataset containing flood and non-flood images from various districts of Sindh.

2.2. Data Preprocessing

Image resizing: Before feeding data to the individual and deep ensemble model, all the images were resized in order to ensure a fixed size input size. The deep learning models accept a fixed-size input because the elements of their architecture do not generalize on the different size of input data.

Normalization The pixel values of the flood and non-flood images were normalized to scale them in a standard and consistent range, usually between 0 and 1, which enhances model convergence during the training by preventing large gradients and ensuring numerical stability. It also reduces the training time of the model.

2.3. Deep Learning Model Training

After performing preprocessing on each dataset, four different deep learning models were trained. Each model was designed to be more complex than the other model. Each model contained a distinct number of layers and parameters, tailored to enhance the representation and extraction of features for confident flood and non-flood image classification. Similar to VGG16 [11], each model was designed in a sequential manner, where each convolutional layer was followed by the max pooling layer. The kernel size in each convolutional layer was set to 3x3. All convolutional layers were followed by ReLU activation function. To avoid the problem of model overfitting, a dropout layer with a rate of 0.5 was used. After the flatten layer, each model used two dense layers to extract more abstract features. The last dense layer was employed with Sigmoid function for performing classification. Each model was trained with a binary cross-entropy loss function and Adam optimizer. Each model was trained for 50 epochs and the batch-size selected was 32.

2.4. Model Evaluation

After completing the training step, each deep learning model was evaluated to measure its predictive capability and reliability for flood and non-flood image classification. Commonly used evaluation metrics for the classification problems such as recall, precision, F1-score, and accuracy were employed to measure the performance of the models on the test dataset [12, 13]. The evaluation metrics help to identify the strengths and weaknesses of each model, as well as their contribution to the ensemble. The results demonstrated the effectiveness of the ensemble approach, where it performed slightly better than the individual models.

2.5. Ensemble Model

Following the training and evaluation of individual models, the outputs of all four models were stacked together to form an ensemble model. This approach combined the strengths of each individual model, leveraging their diverse architectures and feature extraction capabilities for better performance. The ensemble model integrated outputs from all models to make final predictions, which significantly improved the accuracy of flood

and non-flood image classification compared to any standalone model. This stacking method ensured that the ensemble utilized the complementary strengths of its components to deliver superior performance. Furthermore, this shows that the deep learning models when stacked together can enhance the predictive performance.

3. Results and Discussion

The performance evaluation of four different models is presented in Table 1, referred to as Model_1 through Model_4. The models were trained and tested on the relevant two datasets, and their results are shown in Table 1. The results include the number of layers, parameters, training accuracy, and validation accuracy. The training and validation accuracy provide insights into the models' performance and generalization ability. Table 1 summarizes the performance metrics of each model, including the number of layers, parameters, and the training and validation accuracies. It is clear from the table that all models achieved perfect training accuracy (1.0), indicating that the models have successfully learned the patterns in the training data. However, the validation accuracy varies between models, with Model_2 performing the best with a validation accuracy of 0.960, followed by Model_1 with 0.945, and Model_3 and Model_4 both with 0.940 and 0.935, respectively. Model_4, despite having the highest number of convolutional layers (11) and parameters (9826113), achieved a relatively lower validation accuracy of 0.935. Compared to the individual models, the deep ensemble model achieved a training accuracy of 1.0, while the validation accuracy achieved is 0.968. This is a bit better than the individual deep-learning models. The accuracy can further be improved by adding more data and fine-tunning the hyperparameters of ensemble learning model.

Table 1. Model Performance

Model	Layers	Parameters	Train Acc	Val Acc	Precision	Recall	F1 Score
Model_1	6	50467969	1.0	0.945	0.956	0.927	0.941
Model_2	8	23907521	1.0	0.960	0.968	0.947	0.957
Model_3	10	11169089	1.0	0.940	0.977	0.895	0.934
Model_4	11	9826113	0.97	0.935	0.936	0.921	0.933
Ensemble Model	-	-	1.0	0.968	0.959	0.968	0.968

3.1. Training vs. Validation Accuracy

Figure 2 illustrates the training and validation accuracy curves for each model. The curves allow us to visually assess the models' learning behavior during training. While the training accuracy reaches 100% for all models, the vali- dation accuracy reveals the models' ability to generalize to unseen data. Models with a significant gap between training and validation accuracy may be over- fitting, but in this case, the differences are relatively small, indicating that the models are not severely overfitting and are performing well on the validation set. Baseline Comparisons: The paper primarily focuses on deep learning architecture and does not compare results against traditional machine learning models or other existing flood classification frameworks. Including such baselines (e.g., Random Forest, SVM, logistic regression, or simple CNNs).

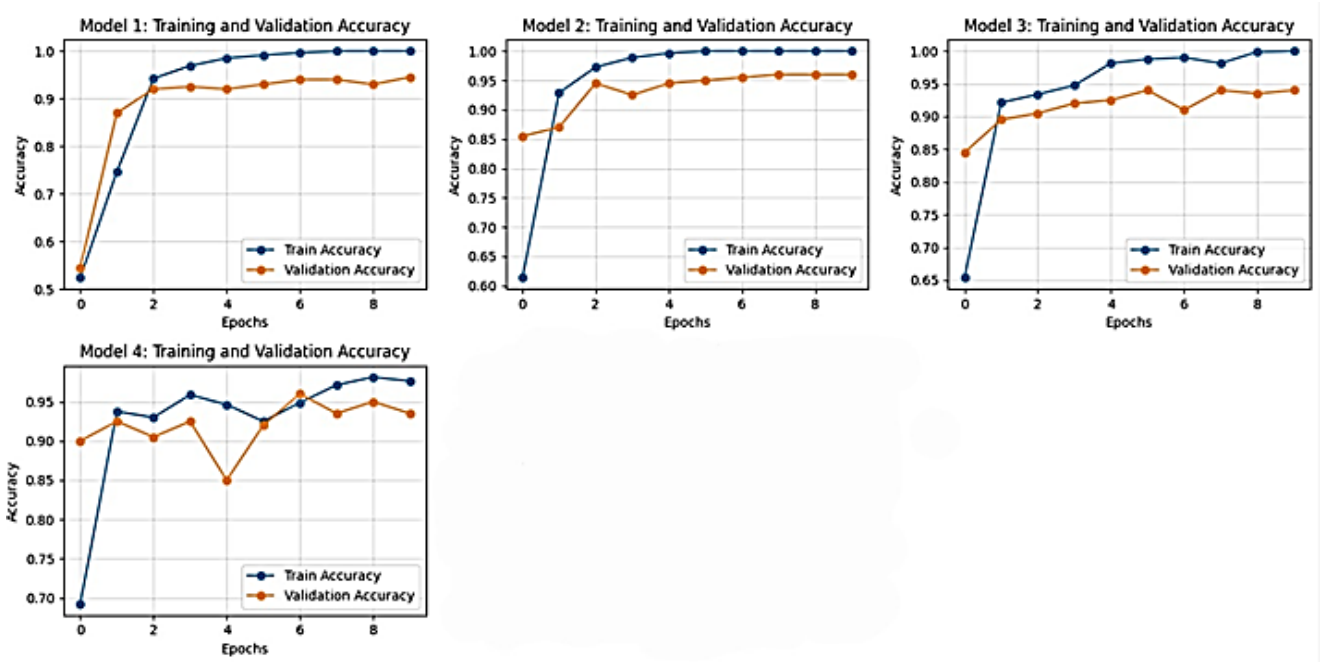

Figure 2. Training and Validation Visualisation of each model

3.2. Confusion Matrix Analysis

The confusion matrices for each model are illustrated in Figure. 3. A confusion matrix is a valuable tool for analyzing how well the model classifies different classes. The diagonal values in the matrix represent the number of correct predictions, while the off-diagonal values represent the misclassifications. Upon reviewing the confusion matrices, Models_1, Model_2, and Model_3 perform well in terms of correct classifications, with few misclassifications. However, model_4, despite its relatively lower validation accuracy, still shows decent performance with some misclassifications in specific categories. Figure 4 illustrates the confusion matrix of a deep ensemble-learning model.

The results indicate that deep ensemble model outperforms the other individual models with the highest validation accuracy of 0.968. This suggests that a smaller architecture with fewer layers and parameters when stacked together can sometimes yield better generalization, possibly due to better regularization or a more suitable model for the dataset. Model_2 also performed well with a validation accuracy of 0.960, indicating that increasing the number of layers (8) and parameters (23,907,521) might lead to a slight improvement in model performance, but it is not as significant as in ensemble model. On the other hand, Models_3 and Model_4, despite having higher numbers of layers and parameters, did not perform as well as expected. This might suggest that increasing the model complexity does not always lead to better performance, and may even lead to overfitting, as seen in Model_4's lower validation accuracy. In conclusion, while the training accuracy was consistently high across all models, the validation accuracy and confusion matrix analysis suggest that ensemble model provides the best balance of learning and generalization. Future work could explore further optimization techniques, such as hyperparameter tuning and regularization, to enhance the performance of the other models.

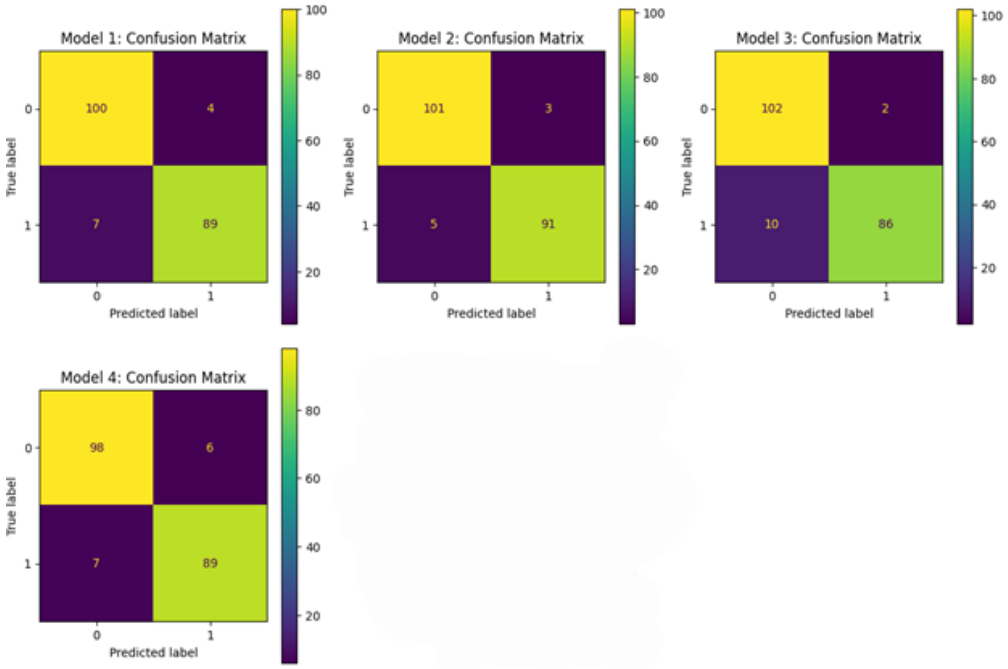

Figure 3. Confusion Matrices of individual deep learning model

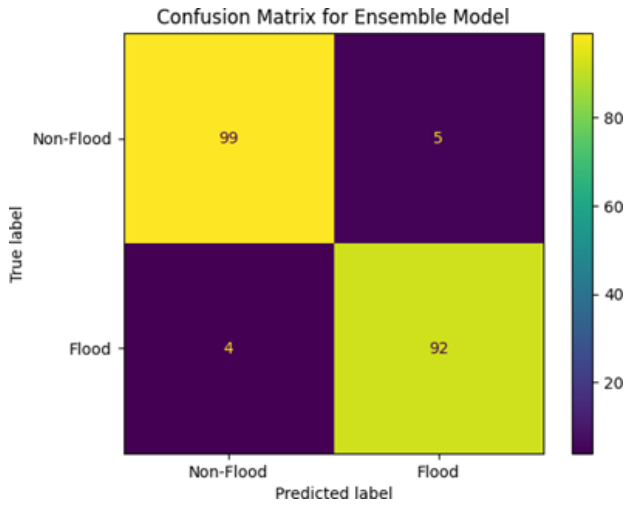

Figure 4. Confusion Matrices of Deep Ensemble Model

4. Conclusion and Future Work

This research work successfully developed a deep learning-based ensemble model for flood and non-flood classification using publicly available datasets. The ensemble approach demonstrated significant improvements in predictive accuracy and robustness over individual deep learning models. By leveraging the diverse capabilities of four deep learning architectures, the framework provided a re- liable solution for identifying flood and non-flood affected areas from images. Future work can explore expanding the dataset to include more diverse environ- mental conditions such as collecting and creating a large-scale dataset of flood and non-flood images from the Sindh Province or other areas of Pakistan. Additionally, integrating real-time classification capabilities and deploying the model in disaster management systems could provide timely and actionable insights for mitigating flood impacts. More advanced deep learning techniques such as attention models and vision transformers can be used to further improve the accuracy of flood and non-flood classification.

Ethical and Practical Implications: There is limited discussion on ethical considerations such as the privacy of individuals captured in flood images, particularly when sourced from social media or public platforms. Future research must ensure compliance with ethical guidelines related to data collection and use, including consent and anonymization. Moreover, practical challenges such as data transmission, real-time inference, and integration with emergency alert systems should be studied to facilitate deployment in real-world flood response systems.

Model Complexity and Real-Time Feasibility: Although the ensemble approach demonstrated superior performance, it inherently increases computational cost due to the stacking of multiple deep models. This complexity may hinder real-time deployment in disaster response systems, especially in resource- constrained environments. Future work should investigate lightweight ensemble strategies or pruning techniques to reduce computational overhead without com- promising classification accuracy. Additionally, benchmarking the model on edge devices or low-latency platforms will help assess practical feasibility.

Conflicts of Interest: The authors declare no conflicts of interest

References

1. Yasi, E., Shakib, T. U., Sharmin, N., & Rizu, T. H.: Flood and Non-Flood Image Classification using Deep Ensemble Learning. Water Resources Management, 1-18 (2024).
2. Bentivoglio, R., Isufi, E., Jonkman, S. N., & Taormina, R.: Deep learning methods for flood mapping: a review of existing applications and future research directions. Hydrology and Earth System Sciences Discussions, 2022, 1-50 (2022).
3. Fayaz, M., Dang, L. M., & Moon, H.: Enhancing land cover classification via deep ensemble network. Knowledge-Based Systems, 305, 112611 (2024).
4. Ullah, F., Ullah, I., Khan, R. U., Khan, S., Khan, K., & Pau, G.: Conventional to deep ensemble methods for hyperspectral image classification: A comprehensive survey. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, (2024).
5. Reis, H. C., & Turk, V.: Integrated deep learning and ensemble learning model for deep feature-based wheat disease detection. Microchemical Journal, 197, 109790 (2024).
6. Leghari, M., Memon, S., Dhomeja, L. D., Jalbani, A. H., & Chandio, A. A.: Deep feature fusion of fingerprint and online signature for multimodal biometrics. Com- puters, 10 (2), 21 (2021.
7. Leghari, M., Memon, S., Dhomeja, L. D., & Jalbani, A. H.: Analyzing the effects of data augmentation on single and multimodal biometrics. Mehran University Research Journal Of Engineering & Technology, 39(3), 647-656 (2020).
8. Kaggle Homepage, https://www.kaggle.com/datasets/hhrclemson/flooding-image- dataset, last accessed 2024/11/30
9. Kaggle Homepage, https://www.kaggle.com/datasets/bharatkaurav/chf2015, last accessed 2024/11/30
10. Salvi, M., Acharya, U. R., Molinari, F., & Meiburger, K. M.: The impact of pre-and post-image processing techniques on deep learning frameworks: A comprehensive re- view for digital pathology image analysis. Computers in Biology and Medicine, 128, 104129 (2021).
11. Simonyan, K.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014).
12. Mosavi, A., Ozturk, P., & Chau, K. W.: Flood prediction using machine learning models: Literature review. Water, 10(11), 1536 (2018).
13. Munawar, H. S., Hammad, A. W., & Waller, S. T.: A review on flood management technologies related to image processing and machine learning. Automation in Construction, 132, 103916 (2021).