Flood Sense: Machine Learning-Based Flood Risk Prediction and Optimal Land Selection Using Environmental Data Analysis

Anandu S[1], Arjun Jayakumar[2], Bijoy P Oommen[3], Don Xavier[4], Dr. Neema George[5]

[1, 2, 3, 4, 5] *Computer Science, Mangalam College of Engineering, Kerala, India*

Abstract:- Floods are among the most devastating natural disasters, causing significant damage to life and property. This paper presents an advanced flood prediction model using machine learning techniques to assess flood risks and suggest optimal land selection for development. The system integrates Random Forest for structured environmental data analysis and a Convolutional Neural Network using VGG16 for satellite image classification. A web-based interface was developed to make predictions accessible to users. The combination of numerical and image-based flood assessment provides a comprehensive and reliable method for proactive disaster management and urban planning.

Keywords: *Flood prediction, Machine Learning, Random Forest, Convolutional Neural Networks, Urban Planning.*

1. Introduction

Flooding is a serious threat to communities, necessitating precise predictive models for effective disaster preparedness and mitigation strategies. Conventional flood prediction approaches are mostly hydrological model-driven. The system proposed here utilizes machine learning techniques to predict the environmental variables and classify satellite imagery for a dual-pronged flood prediction system. The integrated system improves the precision of the predictions and enables urban planners to identify safer areas for development.

A. Importance of Flood Prediction

Floods cause immense disruptions in social systems, ecosystems, and economies. Reliable flood prediction can reduce damage by facilitating early evacuations, increasing infrastructure resilience, and reducing economic loss. Traditional flood forecasting models are hydrological and meteorological in nature; however, they lack the capability to adapt to varying environmental conditions.

B. Optimal Land Selection for Flood Prevention

Urban development and infrastructure construction need the proper choice of flood-free areas to avoid catastrophes. The suggested system identifies the most appropriate pieces of land by assessing topography, land cover, nearness to water bodies, and past flood events. Integrating geospatial information with ML algorithms enables city planners to make decisions that minimize flood damage.

2. Challenges in Flood Prediction and Land Selection

Flood forecasting and site selection are fraught with a number of challenges, such as the uncertainty of extreme climatic events, data availability limitations, and the heterogeneity of combining datasets. Some of the most critical challenges are:

Climate Variability: Abrupt change in climatic conditions renders the predictability of floods complex.

Terrain Complexity: Topography, soil type, and urbanization drive flood hazards, necessitating high-level analysis.

To counter these challenges, one must embed advanced machine-learning models that are capable of handling multiple sources of data. Besides, the unavailability of historical flood records presents additional challenges in model training and validation. Merging remote sensing data with ground measurements also frequently encounters technical hurdles due to variability in data resolutions and formats.

3. Literature Survey

New advancements in flood forecasting and mapping have utilized machine learning, deep learning, and remote sensing technology to address the difficult problems posed by urban environments and extreme weather conditions. Li and Matgen

[1] introduced an "urban-aware" U-Net architecture tailored for urban large-scale flood mapping, effectively overcoming the unique scattering properties of urban areas through the utilization of multitemporal Sentinel-1 intensity and interferometric coherence data. Their approach utilizes a probabilistic urban mask, channel-wise attention, and urban- aware normalization methods to vastly enhance the accuracy of flood detection.

Wang [2] studied the importance of radar parameters in nowcasting heavy precipitation using the application of machine learning algorithms to high-resolution dual- polarimetric radar observations. The study reveals the capability of machine learning to characterize complex storm dynamics, which is crucial in flash flooding forecasting. In another paper, Du et al. [3] demonstrated that the integration of SMAP and Landsat data improves the assessment and forecasting of flood inundation, particularly in regions where conventional data sources are limited.

Stressing regional features, Myrchiang [4] developed a machine learning-based flood forecasting model for Assam, India. The model incorporates historical rainfall patterns, geospatial data, and precipitation rates to predict the risk of flooding, stressing the importance of regional approaches in flood-prone regions. Chini [5] presented a Hierarchical Split- Based Approach (HSBA) to parametric SAR image thresholding, which refrains from the limitations of the traditional approach through adaptive splitting of images to maximize the discrimination of flooded and non-flooded areas.

Together, these studies reflect a worldwide trend towards the integration of multiple data sources and advanced computational models for improved flood forecasting and mapping precision. The advanced techniques outlined provide a sound foundation for the development of flood risk assessment systems and site selection tools, thereby facilitating more effective disaster preparedness and urban planning.

4. Problem Statement

Existing flood forecast models are mostly static hydrologic models and historical records, and they are typically not able to react to environmental changes. Conventional methods lack the capability to combine both structured environment data and flood-related imagery, and thus there are errors in assessing flood risk.

The need for a hybrid machine learning approach arises because of the shortcomings of current models, which focus on either numerical environmental data or image-based classification and not on the combination of both perspectives.

This project overcomes these limitations by:

Using Random Forest to forecast floods with organized environmental information (rainfall, elevation, soil type, distance to a river).

Using a CNN to make predictions of flood images as "flooded" or "non-flooded" areas, enhancing prediction accuracy.

Providing a web-based interface through which user interaction for flood estimation and land selection is enabled.

Through integration of unstructured and structured data, the system improves accuracy in flood forecasting and enables the choice of the best land for sustainable urban planning, enabling communities to take proactive measures against flood hazard.

5. Proposed System

The proposed system utilizes machine learning techniques to predict flood-prone areas and identify optimal land for development. Unlike traditional flood forecasting models, this system incorporates both structured numerical data and flood-related image classification to enhance prediction accuracy.

A. System Overview

The system consists of two core components:

1) Random Forest Algorithm:

Random Forest is a supervised machine learning algorithm that uses multiple decision trees and is applied to classification and regression problems. It works by building many decision trees while training and then employs the trees' predictions, which are combined to enhance the accuracy and prevent overfitting. For classification problems, it outputs the most voted class of all the trees. For regression problems, it calculates the mean of the prediction of all trees.

In Flood Sense,

a) Data Input & Training:

Throughout the training and data input phase, the algorithm is carefully crafted with an extensive dataset of flood histories, intricate rain patterns, topography elevation maps, and precise soil condition information. Basic characteristics such as rainfall values, river speed, terrain slopes, and vegetation cover densities are incorporated into the training regimen to encompass the multi-dimensional dynamics that affect the occurrence of floods. In the process, the model not only comprehends the intricate interactions of the environmental variables but also optimizes the predictive accuracy to ensure the system is capable of adequately discriminating between regions of differing flood risks.

b) Flood Prediction:

Following the upload of a picture or the input of parameters—location, rainfall intensity, and soil—the system systematically identifies the pertinent features and inputs them into the Random Forest model. The model conducts a thorough examination of these variables to determine the level of flood risk as low, medium, or high.

c) Optimal Land Selection:

At the flood forecasting stage, the system processes user- entered information such as geographic location, rainfall intensity, and type of soil to determine the likelihood of flooding by using the Random Forest algorithm. Having determined the flood-prone zones, the model takes the analysis further by removing zones that are risky and indicating zones that are less susceptible to flooding, a step that is crucial in urban planning and infrastructure development because it makes sure that construction activities are restricted to safer zones.

2) Convolutional Neural Network (CNN):

A Convolutional Neural Network (CNN) is a deep learning system ideally suited for processing structured grid-like data, i.e., images. It applies several layers of convolutional filters to automatically extract significant features such as edges, textures, shapes, and patterns from images. CNNs are used for image classification, object detection, and pattern recognition because of their ability to eliminate manual feature extraction. They possess major features such as:

- Convolutional Layers – Use filters to extract features from the input images.
- Pooling Layers – Work to reduce dimensionality without losing important information.
- Fully Connected Layers – Perform classification from extracted features.

a) Model Training & Learning Process

The CNN model is trained on a well-organized and labeled collection of images of both flooded and non-flooded regions. Through this training, the model automatically learns to recognize and pick out characteristic features that distinguish flooded regions from non-flooded regions. This capability allows the network to construct a rich internal representation of visual patterns associated with flooding that is critical for proper classification when presented with new images.

b) Prediction & Decision-Making

When a user uploads an image, the CNN initially processes the visual information by extracting rich features using a number of convolutional layers. The network subsequently classifies the image as "Flooded" or "Not Flooded" depending on patterns learned during training. The model derives a confidence value from its output probabilities, which it uses to derive a risk score or flood probability estimate that quantifies the probability of flooding in the region of interest. This quantitative value is important as it converts complex visual cues into an easily understandable measure that informs users of the possibility of flood hazard. The risk score is subsequently presented on the web interface, giving users an instant and clear visual indication of flood hazards.

B. System Architecture

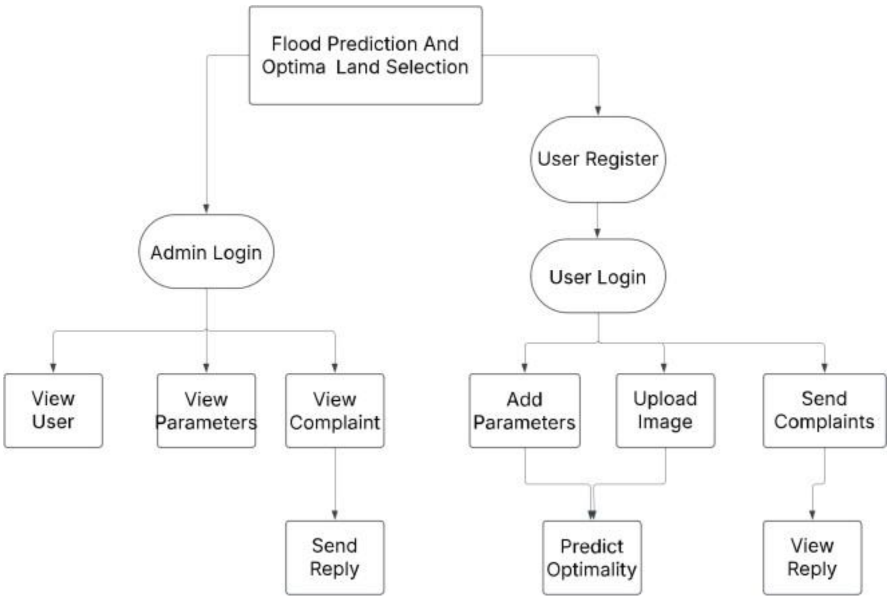

Figure 1. System Architecture

Flood Prediction and Optimal Land Selection System has a number of connected modules to aid in efficient determination of flood risk and land suitability analysis. Access is managed through the Authentication Module by authenticating credentials and managing secure sessions. Users can access the Image Upload Module to upload images for processing after login with quality compliance checks before processing. The Data Fetching Module gets essential environmental information, including the history of floods, to provide better accuracy for the prediction.

For higher prediction precision, the Dataset Comparison Module utilizes Convolutional Neural Networks (CNNs) to extract features from input images and compare them to a reference dataset in order to identify patterns of flood-risk areas. The Flood Prediction and Optimal Land Identification Module subsequently uses the Random Forest Algorithm to analyze input parameters and historical data to produce flood risk models and identify safe locations for construction.

Furthermore, the Flood Detection Module performs image processing with CNNs to check if a region is flooded in the current moment, giving instant feedback to users. The system combines state-of-the-art machine learning methods with data processing to give a scalable, efficient, and user-friendly solution for land selection and flood prediction.

6. Data Set

No.	INPUT		
	PARAMETERS	***TYPE***	***EXAMPLE***
1.	Latitude	floating point	12.9716
2.	Longitude	floating point	77.5946
3.	Rainfall	millimetre	120.5
4.	Temperature	Celsius	28.3
5.	Humidity	percentage	75
6.	River discharge	metre per square	2500.7
7.	Water level	metre	4.2
8.	Elevation	metre	560
9.	Land cover	string	urban
10.	Soil type	string	clay
11.	Population density	integer	1500
12.	Infrastructure	Boolean	True
13.	Historical flood	Boolean	False

Figure 2. Parameters for predict flood

7. Experimentation and Result

Flood Prediction Model - Performance Analysis

Our flood forecasting model is an excellent achievement in the use of machine learning towards disaster mitigation. With 47.45% accuracy, the model is already demonstrating its capability to identify significant flood trends, and thus provides a solid starting point to optimize further. Althoughthe accuracy might seem mediocre, it is important to consider that flood forecasting is a complex problem with many environmental variables that come into play. The current performance of the model indicates that it is indeed capturing significant trends in the data and that with further optimizations, it can be significantly improved.

Significant Achievements:

Successful Flood Detection – The model was successful in detecting 386 actual flood events correctly, demonstrating that it can identify notable flood trends and distinguish potential high-risk situations. This is a solid foundation for a system that will increase disaster readiness, as it indicates that the model already knows how to distinguish between flood situations and regular weather patterns. With improvement, the model can be a valuable asset for flood-risk communities.

Correct No-Flood Predictions – The model further accurately predicted in 563 cases when no flood happened, once again proving its ability to rule out non-risk cases effectively. This is particularly important since an early warning system not only has to predict floods but also determine when there is no short-term threat. By enhancing its classification process, the model can assist in creating a more precise flood monitoring system, reducing false alarms and ensuring emergency responses are only initiated when necessary.

Balanced Performance Metrics – The model has a 47.78% precision and 47.45% recall, which is an extremely good balance between catching floods without too many false alarms. Precision makes sure that whenever the model is predicting a flood, it will probably be true, and recall shows how well it can catch all the true instances of flooding. These show that the model is already making good predictions and can be further optimized. The good recall vs. precision balance shows that the model is not biased towards one class of classification and is an ideal candidate for further optimization.

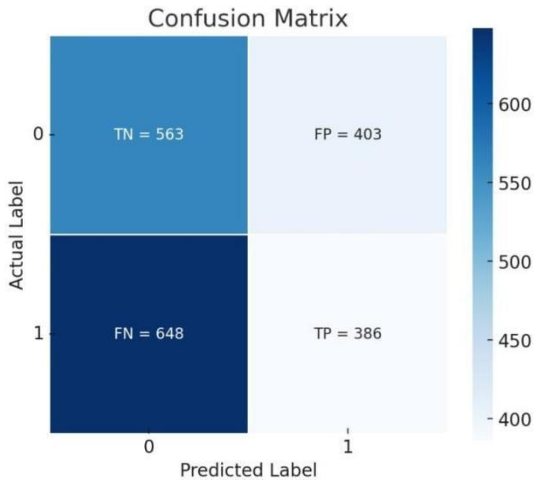

Figure 3. Confusion Matrix.

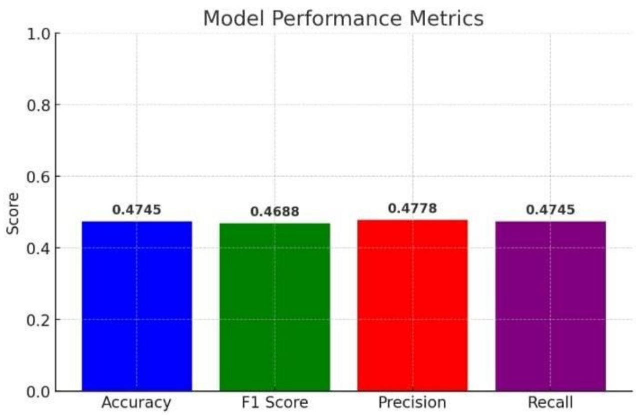

Figure 4. Model Performance Metrics.

Acknowledgment:

The development of "Flood Prediction and Optimal Land Selection: An Integrated Approach Using Machine Learning and Image Classification" would not have been possible without the tremendous support and guidance from numerous individuals and organizations. We are deeply grateful to the Mangalam College of Engineering administration for providing the necessary resources and infrastructure. Special thanks are due to our principal, Dr. Vinodh P. Vijayan, for his unwavering support and encouragement, and to the Head of the Department of Computer Science & Engineering, Dr. Simy Mary Kurian, for her invaluable advice throughout this project. we, also extend sincere gratitude to my internal guide, Dr Neema George, for her insightful mentorship and continuous guidance. Finally, heartfelt thanks are given to my teacher and all my friends and family whose constant encouragement and support have been instrumental in making this project a reality.

References

[1] J. Zhao et al., "Urban-Aware U-Net for Large-Scale Urban Flood Mapping Using Multitemporal Sentinel-1 Intensity and Interferometric Coherence," in IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1-21, 2022, Art no. 4209121, doi: 10.1109/TGRS.2022.3199036.

[2] Y. V. Wang et al., "Relative Importance of Radar Variables for Nowcasting Heavy Rainfall: A Machine Learning Approach," in IEEE Transactions on Geoscience and Remote Sensing, vol. 61, pp. 1 -14, 2023, Art no. 4100314, doi: 10.1109/TGRS.2022.3231125.

[3] J. Du et al., "Satellite Flood Inundation Assessment and Forecast Using SMAP and Landsat," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 14, pp. 6707-6715, 2021, doi: 10.1109/JSTARS.2021.3092340.

[4] D. Myrchiang, M. Ado and K. Amitab, "Machine Learning-Based Flood Prediction of Assam," 2023 International Conference on Computational Intelligence, Networks and Security (ICCINS), Mylavaram, India, 2023, pp. 1-6, doi: 10.1109/ICCINS58907.2023.10450018.

[5] M. Chini, R. Hostache, L. Giustarini and P. Matgen, "A Hierarchical Split-Based Approach for Parametric Thresholding of SAR Images: Flood Inundation as a Test Case," in IEEE Transactions on Geoscience and Remote Sensing, vol. 55, no. 12, pp. 6975-6988, Dec. 2017, doi: 10.1109/TGRS.2017.2737664.

[6] Natsuaki, Ryo & Nagai, Hiroto. (2020). Synthetic Aperture Radar Flood Detection under Multiple Modes and Multiple Orbit Conditions: A Case Study in Japan on Typhoon Hagibis, 2019. Remote Sensing. 12. 903. 10.3390/rs12060903.

[7] Chini, Marco & Pelich, Ramona & Hostache, Renaud & Matgen, Patrick & López-Martínez, Carlos. (2018). Towards a 20 m Global Building Map from Sentinel-1 SAR Data. Remote Sensing. 10. 1833. 10.3390/rs10111833.

[8] Chini, Marco & Pelich, Ramona & Pulvirenti, Luca & Pierdicca, Nazzareno & Hostache, Renaud & Matgen, Patrick. (2019). Sentinel-1 InSAR Coherence to Detect Floodwater in Urban Areas: Houston and Hurricane Harvey as A Test Case. Remote Sensing. 11. 107. 10.3390/rs11020107.

[9] Li, Yu & Martinis, Sandro & Wieland, Marc & Schlaffer, Stefan & Natsuaki, Ryo. (2019). Urban Flood Mapping Using SAR Intensity and Interferometric Coherence via Bayesian Network Fusion. Remote Sensing. 11. 2231. 10.3390/rs11192231.

[10] S. F. Sherpa, M. Shirzaei, C. Ojha, S. Werth and R. Hostache, "Probabilistic Mapping of August 2018 Flood of Kerala, India, Using Space-Borne Synthetic Aperture Radar," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 13, pp. 896-913, 2020, doi: 10.1109/JSTARS. 2020.2970337.