Journal of Engineering, Science and Technological Trends (JESTT)

Vol. 2, Issue 3, November 2025
DOI: https://doi.org/10.64060/jestt.v2i3.2

Review

# Natural Disaster Prediction and Mitigation through Machine Learning

Fatima Abbas[1], Jamil Afzal[2*] and Sidra Shahid[1]

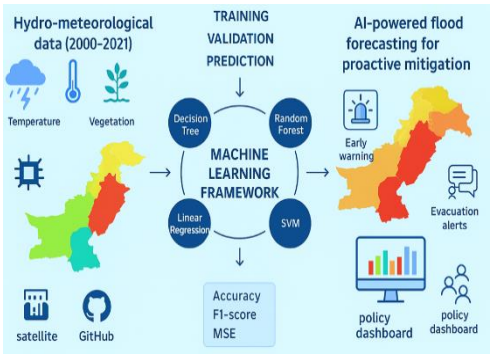
[1]University of Agriculture Faisalabad, Pakistan
[2]International Islamic University, Kuala Lumpur, Malaysia
* Corresponding Email: jamilafzal@iium.edu.my (J. Afzal)

Received: 06 September 2025 / Revised: 13 October 2025 / Accepted: 09 October 2025 / Published online: 15 November 2025



## ABSTRACT

Flooding remains a major natural disaster affecting Pakistan's provinces of Punjab, Sindh, Khyber Pakhtunkhwa, and Balochistan, with increasing severity due to climate change and human activities. This research explores the application of machine learning techniques to enhance flood prediction accuracy for the years 2025 to 2030. The study utilises historical hydro-meteorological data, including rainfall, temperature, and vegetation indices, to train four machine learning models: Decision Tree, Random Forest, Linear Regression, and Support Vector Machine (SVM). Standard evaluation metrics such as precision, recall, F1-score, and mean squared error (MSE) are used to assess model performance. Results show that Random Forest and SVM outperform the other models in terms of both accuracy and generalizability. These models effectively identify high-risk flood zones across the studied provinces. The findings demonstrate the potential of data-driven approaches to support early warning systems, enabling better disaster preparedness, resource allocation, and mitigation planning. This research highlights how machine learning can play a critical role in reducing flood-related risks and enhancing resilience against future natural disasters in Pakistan.




**Keywords***:* Climate Change; Disaster management; Flood Forecasting; Machine learning

## 1. Introduction

Natural disasters, particularly floods, continue to pose significant threats to human lives, infrastructure, and economies across the globe. The increasing frequency and intensity of such disasters over the past few decades can largely be attributed to climate change, rapid urbanisation, and ongoing environmental degradation [1]. Among these disasters, floods rank as one of the most common and devastating, affecting millions of people annually. They disrupt transportation, communication, and supply chains, while also displacing communities and causing long-term financial losses [2].In response to this growing threat, accurate and timely flood prediction has become critical in minimising the risk to both lives and livelihoods. The evolution of modern technologies, particularly in the domains of data science and machine learning (ML), has opened promising avenues for developing more accurate and adaptive flood forecasting systems. These technologies can process complex and large-scale environmental datasets to identify hidden patterns, offering faster and more informed decision-making for disaster management [3].

Previous research efforts have primarily focused on traditional hydrological models and early-stage statistical techniques for flood prediction. Commonly used models include Rainfall-Runoff models, GIS-based simulation tools, and various statistical forecasting methods, which have provided important baseline insights into flood [4],[5]. However, these approaches often struggle to adapt to real-time conditions and lack the flexibility required to integrate changing environmental variables dynamically. In recent years, the focus has shifted towards machine learning techniques due to their capacity to analyse non-linear relationships and improve predictive performance. Researchers such as Brunner et al. (2021) have demonstrated how models like Random Forests, Decision Trees, Artificial Neural Networks, and Support Vector Machines can offer improved forecasting capabilities [6]. However, many of these efforts have been limited by static or incomplete datasets, insufficient

geographical coverage, or a lack of model validation in real-world disaster scenarios. Data scarcity, limited sensor availability, and inconsistent environmental monitoring especially challenge flood prediction efforts in developing countries like Pakistan [7],[8].

Addressing these limitations, the current study proposes a robust machine learning-based flood prediction framework specifically designed for Pakistan's four most vulnerable provinces: Punjab, Sindh, Khyber Pakhtunkhwa, and Balochistan. Historical flood-related data from 2000 to 2021 have been utilised, focusing on essential environmental variables such as rainfall, temperature, vegetation index, and ice melt levels. The selected machine learning models, Decision Tree, Random Forest, Linear Regression, and Support Vector Machine (SVM), were trained and validated using standard benchmarks to ensure reliable and generalised performance. Monthly flood risk forecasts were generated for the years 2025 to 2030, providing an extended window for future disaster planning. The results are intended to support government agencies, policymakers, and emergency response teams in developing region-specific early warning systems and mitigation strategies [9].

The novelty of this study lies in the following contributions:

1. To focus on region-specific flood forecasting using machine learning, tailored to Pakistan's diverse climate zones and historical flood data.
2. To find the most accurate and generalizable ML model by comparing multiple algorithms under consistent validation benchmarks.
3. To explore the impact of hydro-meteorological variables - including temperature, rainfall, vegetation index, and ice - in enhancing flood prediction accuracy.
4. To use ML-generated monthly predictions (2025–2030) for aiding disaster management authorities in early warning systems and risk reduction planning.

## 2. Prior Investigation

### 2.1. Machine Learning in Flood Prediction: An Overview

The implementation of Machine learning models alongside classic statistical techniques has greatly increased the speed and reliability of flood predictions. Merely a few years ago, workers in meteorology and environmental science were incorporating machine learning in their forecasts. Earlier, hydrologists would always try to anticipate flood periods with the help of statistical models based on certain historic events. While these models have been useful in certain scenarios, they fail to account for multi-dimensional non-linear outcomes, rendering them ineffective in rapidly changing climatic conditions. The application of ML in these cases has remedied the deficiencies of earlier models. Rather than having to make guesses, ML can now fully extract complex relationships from large datasets and make precise predictions [10]. Originally, hydrological prediction models were expected to be a lot more accurate with the integration of other ML techniques. The reasoning behind the incorporation of physical parameters in earlier models is to strive for the maximum attainable results, which is not the case in machine learning [11]. Among the many methods in ML, the most widely utilised ones are Decision Trees and Random Forests due to their efficiency in high-dimensional datasets. These models also have some modifications which allow the analysis of both categorical and continuous variables at the same time, deal with the missing data optimally, and capture the model parameters by post-estimation processing. This helps in the assessment and control of risks related to potential flood damage [12].

Alongside decision trees and random forests, they have examined the application of more advanced ML methods like Support Vector Machines (SVM) and Artificial Neural Networks (ANN) in the area of flood prediction. SVM is particularly strong in performing classifications of meteorological data because it can simultaneously cluster a multitude of interactions from different flood indicators and also overfit them. At the same time, ANN mimics the functions of a human brain, which allows the model to capture complex dependencies and non-linear relationships among different sets of environmental data. These particular features of SVM and ANN models enable the construction of predictive models of flood events in areas with rapid climatic changes and where most models fail [13]. Implementing machine learning models to flood forecasting has greatly enhanced predicting and making provisions for disasters in numerous flood-prone areas, which is remarkable. These areas have benefited from these models as it helped improve disaster preparedness. With each new feature, a significant amount of information is received; thus, the possible enhancement of machine learning models to assist in flood risk management for decision makers, responders, communities, and others to anticipate and minimise flood impacts is immense. The study of the area requires an urgent application of sophisticated deep learning techniques and hybrid models that will aid in more accurate and effective flood forecasting owing to climate change and other environmental factors.

### 2.2. Effectiveness of Ensemble Learning Techniques

An ML technology for flood prediction has one of the greatest benefits in the application of ensemble learning strategies, which, through the use of multiple algorithms, have fundamentally transformed predictive models by enhancing their accuracy and effectiveness. Single-model approaches, in some cases, may be helpful, but they are often associated with overfitting, high variance, and sensitivity to noise in the data. On the other hand, ensemble approaches combining the predictive capabilities of numerous models mitigate the mentioned constraints and provide improved generalisation as well as robustness. Empirical evidence has shown that ensemble learning methods, specifically Random Forest, consistently outperform the use of conventional single-model methods by lowering their variance and overfitting, and in turn improve the stability and reliability of flood predictions [14]. This stability in particular, is greatly important in flood forecasting because the accuracy of the prediction has potentially severe consequences within disaster mitigation and preparedness regions. Random Forest is one of the most popular methods of ensemble learning. It is an extension of Decision Trees that improves prediction by building several decision trees and combining their predictions. Unlike a Decision Tree, which can easily overfit and become overly responsive to variations within training data, Random Forest builds a super set of trees by deliberately adding randomness within feature and training sample selection. This helps ensure that the model does not depend too much on a single predictor, hence improving generalisation and the chances of avoiding errors due to noisy or incomplete data [15]. As a flood prediction tool, Random Forest is a good candidate due to its strong performance in high-dimensional datasets, because so many variables comprising the precipitation, river discharge, temperature, and land surface change intimately monitor each other. The use of a collection of individual decision trees makes Random Forest easier to use without compromising prediction accuracy, which is ideal for hydrologists and disaster management officials.

Different ensemble learning methods, like boosting algorithms, have also become popular because of their ability to further improve the predictive accuracy of ML-based flood forecasting models. Some of the boosting methods are: Adaptive Boosting (AdaBoost), Gradient Boosting Machines (GBM). These methods assign new weights to weak learners, in turn making them better. For example, AdaBoost gives higher weights to misclassified instances, which forces the following models to focus on correcting

previous mistakes, thus improving the accuracy of flood predictions [16]. Likewise, GBM builds trees in a forward-stage manner, optimising loss functions for each stage to improve the precision of flood-prone area identification. These boosting methods have been able to achieve the identification of the intricate features and the nonlinear dependences of hydrological variables, which is important for estimating the floods. The comparison done between the ensemble learning models and the traditional hydrologic models has shown that the effectiveness of ensemble approaches is better in predicting floods. Some studies show that some classical models, like the popular rainfall-runoff models, for instance, tend to underestimate the variability of meteorological measurements because they operate on predefined relationships between their input factors. However, Rahman et al. argue that ensemble-based ML techniques are more reliable in predicting uncertain environmental conditions [17]. Their ability to utilise multiple predictors and adapt to new data enables them to overcome uncertainties in their environment.

Maintaining flexibility is crucial for areas affected by climate change since the rainfall, extreme weather, and river flow are highly variable. These areas require better models for predicting and controlling flood disasters. The application of ensemble learning approaches in the automation of the processes connected with disaster forecasting has been described in a number of examples, particularly in regions classified as flood regions, where precise prediction is necessary for quick decision-making during a disaster. For instance, Ruichen et al. reported that Pakistan is susceptible and studies proved that ensemble learning models are performing better than the conventional techniques of forecasting, which rely on excessive monsoon rains, glacial thawing, and river flooding [18]. These models have significantly enhanced shallow disaster management in the area by providing timely and accurate predictive warnings, which enable the responsible authorities to evacuate and reinforce the required civic infrastructures. Furthermore, the application of ensemble learning techniques, together with real-time data collection systems such as satellite remote sensing and the Internet of Things (IoT), has made it possible to respond faster and more accurately to flood risks.

### 2.3. The Need for Province-Specific Model Selection

Although machine learning (ML) models for flood predictions have greatly advanced the accuracy of forecasting floods, the accuracy of these models differs greatly across regions due to differences in climate and hydrology. In Pakistan, which has an extremely diverse topography including the towering mountains of Khyber Pakhtunkhwa and Gilgit-Baltistan, fertile plains of Punjab and arid deserts of Balochistan, a single flood prediction model will most likely not provide accurate predictions within all provinces. These interprovincial differences require that ML model choice be tailored to specific provinces in order to achieve favourable results that consider the geophysical and climatic realities of the region [19]. The need for an appropriate selection of ML models in Pakistan is illustrated through the different flood mechanisms across the country. For example, in Punjab and Sindh Phases, floods are mainly attributed to the monsoons and river flooding from major rivers like the Indus, Jhelum, Chenab, Ravi, and Sutlej. With respect to river flooding, it is found that ensemble methods, including Random Forests and Gradient Boosting Machines (GBM), outperform others because they are simpler to use with large, structured datasets and are powerful in merging multiple meteorological and hydrological parameters [20]. Furthermore, these models capture data about rainfall, river water flow, soil moisture, and previous floods to help estimate possible floods. This helps the policymakers to devise plans that could mitigate potential risks and damages. Moreover, these regions are estimated for large magnitude floods because there are available methods for missing value reconstruction, multifactor examination, and large-scale hydrologic database application.

The dynamics of flooding in Khyber Pakhtunkhwa and Balochistan are predominantly governed by the thawing of glaciers, excessive rains from the mountains, steep slopes, and the region's low soil absorption rates. In comparison to Punjab and Sindh, these regions experience much more irregular, exotic, and violent flash floods instead of the more systematic river flooding. Recent investigations have shown how these provinces perform with Support Vector Machines (SVM) and Artificial Neural Networks (ANN) and most have astonishingly good results owing to the ability of these methods to model complex nonlinear interactions of the most important meteorological factors, which include the temperature, local rainfall, and snowmelt [21]. That is crucial for SVM in the situation when a vast number of environmental drivers in these regions strongly interact to form a highly complex classification. Much like ANNs, SVMs have been capable of capturing the nonlinear relationships in the occurrence of flash floods, thus enabling the activation and dynamic simulation of hydrological responses to climatic variations. In light of these differences on a provincial scale, the specialists highlight the need for developing training sets that correspond to the meteorological and hydrological characteristics of every region. Every flood prediction model associated with the implementation of ML techniques performs effectively with the right input data; thus, local climatic conditions, historical flooding, and sensor data are essential for the right predictions [22]. For example, Punjab and Sindh possess rivers whose data must be included in the databases of river inflow, precipitation, and groundwater level changes. Conversely, Khyber Pakhtunkhwa and Balochistan need to focus more on glacier movement, the land surface configuration, and the occurrence of intense storms. Ignoring such parameters would most likely lead to forecasting errors which would defeat the whole purpose of the ML flood prediction systems that disaster management organisations have to work with.

The use of IoT sensors together with remote sensing devices enables better integration of data to enhance the accuracy of province-level ML models in real time [23]. Active monitoring reveals a number of environmental factors, for example, remote sensing of surface water, precipitation, and changes in the land surface, which have the potential to increase the advantage that machine learning models could have if these parameters are provided as real-time inputs [24]. Punjab and Sindh's regions of Punjab and Sindh are served by the Sentinel 1 and Landsat satellites, which help in observing river basin borders and help in advancing flood models by detecting conditions that may cause river overbanking. Enabling real-time flood prediction from glacier melting, local rainfall, and soil moisture IoT-based sensor networks in the mountainous regions of Khyber Pakhtunkhwa and Balochistan greatly assists [25]. These data sources make it possible for ML algorithms to design sophisticated dynamic forecasting systems that increase intervention and disaster management system functionality.

In addition, further work should be devoted to the design of optimal hybrid models that combine several machine learning algorithms for alleviating regional discrepancies in flood prediction precision. It is true that deep learning approaches, or any single machine learning algorithm, will outperform the rest in one environment while underperforming in others, which is why there is a growing interest in hybrid models that combine different approaches. For example, the combination of ensemble learning algorithms with deep learning Convolutional Neural Networks (CNNs) is known to improve the accuracy of flood prediction by more fully exploiting structured and unstructured data patterns [26]. Similarly, other hybrid models that combine physical hydrological models and machine learning prediction models may improve flood fore-

casting and permit broader usage in different regions. The adoption of a provincial-level machine learning model selection is particularly important considering the climate change-induced irregularity of floods in the region of Pakistan. Over the past ten years, there has been an increased incidence of extreme weather, with the addition of monsoon rains being increasingly warmer, warmer temperatures, melting of glaciers, and other more recent changes in rainfall distribution further worsening the flooding's forecasting problem. In order to tackle these problems, there is a need to develop real-time climate impact adaptive ML frameworks that would strengthen resilience to disastrous flooding. It is beyond doubt that there is a requirement to fundamentally rethink the floods' prediction repeating periodicity and also the algorithmic models which should be implemented in the provinces of Pakistan to enhance the preparedness and response to the changing, complex and adverse hydrological threats.

## 3. Methodology

### 3.1. Data Collection

In order to create a credible and precise machine learning (ML)-based flood prediction model, a dataset with all relevant features concerning floods and their contributory aspects was collected from all possible sources, with a focus on publicly available open source datasets from GitHub. The dataset is compiled for the years 2000 to 2021 with the objective of having a sufficient time history of actual floods and the causative factors for floods, like the environment during that time period and to understand the outer phenomena over a longer term, which helps to better capture accurate results to enhance the ML models' predictive scope over time. The dataset was organised from 2000 to 2021 so as to capture insights over extensive durations, which helps in capturing cyclical tendencies. The criteria for research materials were determined after a thorough examination of available study materials and discussions with specialists in hydrology and climate sciences. The important variables selected in the dataset are precipitation, temperature, ice, and vegetation index data. These features were incorporated due to their significant influence on the incidence of floods, which is testified in past records [27]. Floods are caused primarily due to excessive downpour and supporting data explains rainfall as one of the strongest factors, alongside temperature, on the chances of floods and urban flooding [28]. Temperature, along with rainfall, exerts strong impacts, but understanding seasonal temperature variations along with snow, melting and glaciers' retreat provides deep insights into regions of Khyber Pakhtunkhwa and Gilgit-Baltistan, where glacial melting floods are prevalent. Gathering ice level data assisted in tracking the seasonal and long-term shifts in glacier masses since GLOF events have become more common lately because of climate change. The dataset consisted of satellite iceberg level measurements, which help evaluate how much glaciers are melting and the flooding that occurs downstream as a result [29],[30]. Also, deriving from the flood satellite imagery and remote sources, the vegetation index was added to track changes in land cover, soil moisture retention, deforestation, and vegetation growth that impact the frequency and severity of floods. Regions with greater vegetation cover tend to absorb more water and lessen the potential for floods, while deforested areas are prone to increased runoff and flooding because of less vegetation. The dataset was collected from a number of public GitHub repos which collect and process climate datasets from reputable sources such as NASA, NOAA, ESA, PMD, and many others. These repositories provided raw climate data, remote data, and precipitation records and then sanitized the data to make it appropriate for machine learning model training and testing. In the interest of assuring reliable data, a series of preprocessing steps were executed, including data normalization, missing value imputation, and outlier treatment. Historical discrepancies and gaps within the records were corrected by checking them against available data and verifying its correctness. Moreover, archival flood event information was integrated with other datasets to enable the model to better adapt to climate changes over time. The cleaned dataset was also split into training and testing datasets to ease model development and evaluation. Later stages of the study could incorporate more real-time data from IoT devices, which would increase the accuracy of the models.

### 3.2 Data Preprocessing

To ensure the quality and reliability of the dataset before training the machine learning (ML) models, several preprocessing techniques were applied. Proper data preprocessing is essential for minimising biases, handling inconsistencies, and optimising model performance. The primary steps included handling missing values, feature selection, and data normalisation.

***Handling Missing Values***: In order to maintain the dataset's accuracy and relevance prior to training the machine learning (ML) models, a set of preprocessing methods was conducted. Appropriate data categorisation is crucial for reducing biases, correcting errors, and improving the model's effectiveness. These procedures mostly comprised missing data treatment, relevant feature selection, and data standardisation.

***Feature Selection***: A critical stage in data cleansing revolved around pinpointing the parameters that had the greatest influence on flooding. A correlation study was undertaken by applying Pearson's correlation coefficient to determine the associations between independent variables (precipitation, temperature, ice, and vegetation index) and a dependent variable (flooding). Those features which showed a strong correlation with flood occurrence were preserved, while those with very little impact were eliminated to ensure noise reduction and improve model effectiveness. A Variance Inflation Factor analysis was also performed to test for multicollinearity among features to ensure that non-essential variables did not bias model estimate outcomes.

***Data Normalisation***: MinMaxScaler was implemented to scale features to the same range of values. All numerical features were scaled to lie within the range of zero and one. This approach helps preserve the original distributions while preventing features from skewing model training. Such normalisation was particularly effective in enhancing convergence and guaranteeing even weight distribution across predictors. These methods were particularly useful for SVM and Linear Regression, which are sensitive to feature magnitudes.

## 4. Model Selection and Implementation

In order to create a strong and reliable system for flood prediction, four separate machine learning models were implemented with Scikit-learn. These models were selected because they could separately tackle classification, regression, and ensemble learning, which are part of flood forecasting.

***Decision Tree***: Classification based on Gini impurity was structured using a Decision Tree model. This model was picked because of its flexibility and ease of interpretation for non-linear relationships between input variables. Decision Trees break down the dataset into successive layers of decision points, which is helpful in pinpointing the limits of flood-inducing parameters, aiding in flood.

***Random Forest***: In order to refine accuracy and mitigate overfitting, a Random Forest model was applied as an ensemble learning technique. Random Forest enhances the effectiveness and dependability of flood forecasting by constructing multiple Decision Trees and combining their outputs. This approach is especially

beneficial for high-dimensional data and for cases with absent data points while reducing bias.

***Linear Regression***: A linear regression was incorporated to estimate flood severity by using numerical variables like rainfall and river water level. This model attempts to partition the flooding by quantifying it, thereby ascertaining the impending threat level of floods by applying a linear function to this dataset.

***Support Vector Machine (SVM)***: An SVM model differentiated whether a flood occurred or not (binary outcome) by classifying it as a flood or no flood. With this, an SVM model can effectively classify data points based on any historical data. SVMs were selected for this project due to their efficient performance with high-dimensional data and their tendency to overfit smaller datasets. The following Figure 1 represents the average accuracy distribution of machine learning models for flood prediction.

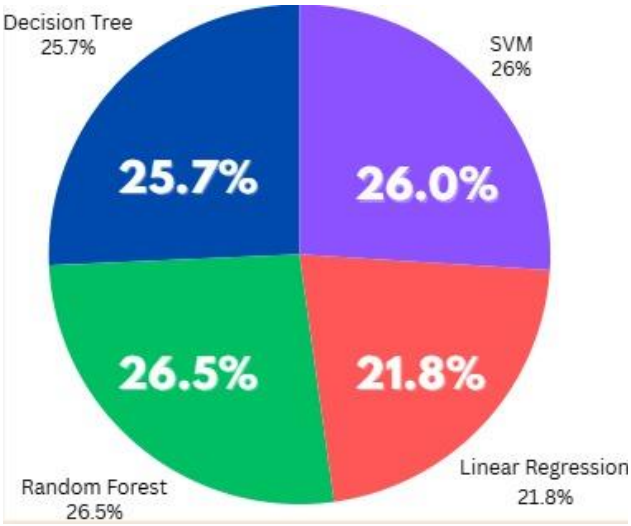



**Figure 1**: *Average accuracy distribution of machine learning models for flood prediction*

To assess the performance of the implemented models, several evaluation metrics were used, ensuring a comprehensive analysis of their predictive accuracy and reliability. The selected metrics included:

- **Accuracy:** Measures the proportion of correct predictions out of total predictions, providing an overall assessment of model effectiveness.
- **Precision:** Evaluates the proportion of true positive flood predictions out of all positive predictions, indicating the model's reliability in correctly identifying flood occurrences.
- **Recall:** Measures the ability of the model to detect actual flood events, crucial for minimising false negatives and ensuring timely warnings.
- **F1-Score:** A harmonic mean of precision and recall, balancing both metrics to provide a robust performance measure, particularly for imbalanced datasets.
- **Mean Squared Error (MSE):** Applied to regression models like Linear Regression, MSE quantifies the average squared difference between actual and predicted flood severity levels, helping assess prediction accuracy in continuous output variables.

Table 1 summarises the model accuracy across the four provinces.

As shown in Figure 2, Random Forest and SVM consistently outperform other models, demonstrating superior predictive capabilities. Figure 3 represents the Actual vs predicted occurrence of floods in Punjab, 2025-2030.

- **Punjab:** High flood risks in July due to peak monsoon rainfall.
- **Sindh:** Seasonal variations with increased risks from June to August.
- **KPK:** Notable risk fluctuations, with high risks in mountainous regions.
- **Balochistan:** Moderate flood risks, with peaks in July-August.

**Table 1**:
Model accuracy across the different provinces of Pakistan

| Province | Decision Tree (%) | Random Forest (%) | Linear Regression (%) | SVM (%) |
|---|---|---|---|---|
| Punjab | 88.13 | 99.15 | 42.23 | 99.15 |
| Sindh | 96.98 | 96.98 | 96.98 | 96.98 |
| KPK | 99.92 | 98.40 | 90.83 | 90.83 |
| Balochistan | 97.17 | 99.00 | 93.50 | 99.00 |

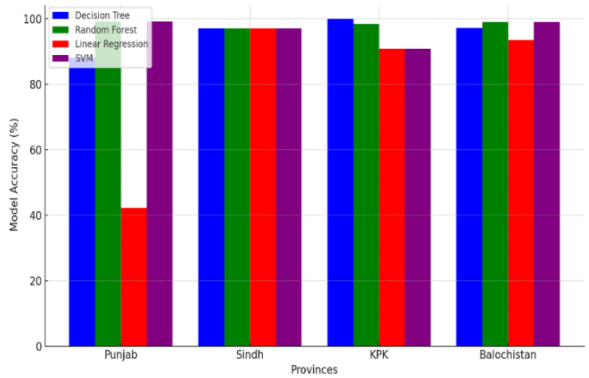



**Figure 2**: *Comparison of Model Accuracy Across Provinces*

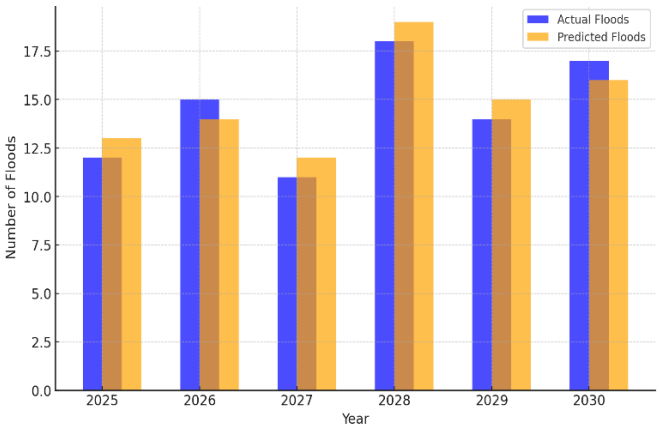



**Figure 3**: *Actual vs Predicted occurrence of flood in Punjab*

## 5. Discussion

This study's findings verify that ensemble models, especially Random Forest, are more effective than other machine learning methods in accurately predicting floods. The effectiveness of Random Forest can be attributed to its capability to handle large and complex multidimensional data sets without suffering from overfitting or data imbalance. Due to the aggregation of multiple decision trees, the model comprehensively captures the complex underlying

relationships among environmental variables and is well-suited for regions prone to flooding that exhibit high variability of climatic and hydrologic phenomena. The model's strength in dealing with absent data, the ability to capture categorical and continuous data, and the diminishment of noise within predictions showcase the model's efficiency in flood forecasting. In this study, one of the major benefits of using Random Forest was the ability to obtain accurate flood forecasts in various regions without changes in rainfall, land surface conditions, and water flow changing accuracy. Unlike conventional hydrological models that depend on fixed equations and preconceived notions, Random Forest takes into consideration the real-time data, which makes it more flexible and responsive in predicting flood occurrences. This feature is specifically useful for nations such as Pakistan, which, owing to its varied topography from the glacial Khyber Pakhtunkhwa region to the flat valley of Sindh and Punjab, has to deal with environmental factors which need to be incorporated in the model.

The research also claimed that Linear Regression, which is very popular in predictive analysis, has profound limitations in areas with high climate variability. Using Linear Regression, it is assumed that the relationships of each pair of variables do not alter over time. It also does not estimate nonlinear dependencies in flood-associated data very accurately. In Balochistan and Khyber Pakhtunkhwa regions, where flash floods occur due to sudden and extreme precipitation or ice cap melting, Linear Regression fails to make the necessary predictions. The model does poorly because of its restrictive assumptions about complex interrelated variables, which deeply reduce its accuracy performance. Thus, it is not suitable to serve as an independent flood prediction model in these areas. The findings of this study reiterate the need to develop machine learning models specific to a region in order to enhance the precision of flood forecasting. Due to the geographical diversity of Pakistan, an all-encompassing model will not work for flood prediction. For example, Random Forest outperformed the other methods in Punjab and Sindh, where monsoonal precipitation and riverine flooding are the main issues, but SVM and ANN are better suited to Khyber Pakhtunkhwa and Balochistan, where flooding is associated with glacial lake outburst and sudden shifts in precipitation. This illustrates the need for predictive models which are focused towards the specific hydro-meteorological characteristics of each province as opposed to basing it on a generalised methodology.

The results also emphasise the importance of localised training datasets for improving model performance. The flood prediction models that used region-specific data, which included historical flood patterns, local weather, and topographical data, outperformed all other models trained on more general datasets by a wide margin. This indicates that model predictions can be further improved by the use of real-time data sources such as satellite and remote sensing images or IoT sensors that can modify environmental variables in real-time. Incorporating such real-time data streams into ML models enables dynamic forecasting, which allows the timely issue of flood alerts and better disaster management response.

## 6. Conclusion

Flood is being classified as an increasing risk to life, infrastructure and economic development, especially in areas like Pakistan, where there is climatological and geographical variation which leads to timely and severe flooding. As the decades have passed, flooding prediction has relied more and more on the standard hydrological models, which were useful but limited as they are not capable of describing the complex, non-linear relationships between different environmental factors. However, with the availability of machine learning (ML), the scope for more advanced tools that can handle enormous data sets, recognise concealed patterns and enhance predictions of flood forecasting has become available to researchers and policymakers. The goals of this study are to demonstrate how different ML algorithms can accurately predict floods in Pakistan, arguing the most pronounced results for prediction using Random Forest and Support Vector Machines (SVM). These Models have proven their ability to manage big data or high-dimensional data sets along with uncertainty in meteorological data. Through these advanced models, the study sponsors AI-enabled flood disaster management programs that may substantially improve early warning systems for disaster prevention and minimise losses associated with floods. However, to provide more accurate and relevant adaptation for these kinds of models, further development is needed regarding real-time weather data integration. The integration of machine learning algorithms has advanced flooding prediction as it shifted research from the use of rigid static methods to more adaptive and multifaceted methods that can model the interrelationships between the variables in question. Among the tested ML models, Random Forest and SVM algorithms outperformed other techniques, such as Decision Trees and Linear Regression, on metrics suited for tele monitoring systems. Predictive flooding forecasting, as well as anticipating the extent of flooding, was highly accurate from using Random Forest, a multi-decision tree ensemble learning method. One of the most important features of Random Forest is, however, mediating missing values and the ability to classify and quantify variables without the need for extensive time consumptive preprocessing. Additionally, Random Forest averts single decision tree model overfitting, which compounds error by averaging predictions, which is another strength that augments its effectiveness in regions such as Punjab and Sindh, where monsoonal flooding and overbank flooding due to the Indus River system are common.

On the contrary, SVM was extremely efficient at classifying flood events given the complex meteorological and hydrological factors. Unlike other modelling techniques that rely on regression analysis and use the linear model as a first assumption, SVM deals with non-linear distributions of data, making it appropriate for Khyber Pakhtunkhwa and Balochistan, which are known to be erratic. These provinces are characterised by sudden flash floods caused by heavy rainfall and rapid glacial melt. SVM performed well with some sets of data due to its ability to separate features corresponding to flood conditions from those related to non-flood situations by optimising the decision boundary. While SVM performed best, other models like Decision Trees and Linear Regression did offer some value, but they weren't as reliable or flexible, which is essential for predicting floods with high accuracy. Decision Trees are computationally simple and cheap when it comes to classifying data, but they do suffer from overfitting when applied to datasets with high variability, when it comes as climate. Regression analysis, especially in the form of Linear Regression, which falls under the umbrella of quantitative analysis where equations are used to depict the relationship between the relevant factors and the occurrence of floods, proved ineffective in places where multiple, highly interacting flood drivers exist, such as terrain slope, soil moisture, and wind direction. In terms of prediction, the earlier floods can be projected, the better prepared society can be for responding to the disaster. Predictive alert systems based on Machine learning models, such as Random Forest and SVM, have the potential to assist governmental agencies as well as citizens in protecting themselves from various forms of floods. By warning users days before actual floods take place, these warning systems can enhance resource deployment for aid services, reduce casualties, and minimise infrastructural damage. In regions of Pakistan which are prone to floods, employing SVM and Random Forest models in alert sys-

tems may prove beneficial. Signals for late evacuation due to poorly forecasted floods resulted in large-scale financial losses during the 2022 floods that affected Sindh and Balochistan. These models would enable authorities to provide timely alerts if proper forecasting tools were in place, resulting in proactive measures being employed. There is a pressing need for AI-centred technology in flood forecasting and management, as well as for investment in holistic approaches that incorporate using such models to enhance national disaster management strategies.

## Declaration

**Competing Interests**: 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.

**Author Contribution Statement**: FA, JA, and SS conceived the idea and designed the research; Analyzed and interpreted the data and wrote the paper.

**Funding Statement**: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

**Consent to Publish**: All authors agree to publish this version of the manuscript in JESTT.

**Data Available Statement**: Data will be made available on request by Fatima Abbas (fatimaabbas3815@gmail.com)

**Clinical trial number**: Not Applicable

**Ethical Approval**: N/A

**Consent to Participate**: N/A.

## References

1. El-Magd, S.A.A., B. Pradhan, and A. Alamri, *Machine learning algorithm for flash flood prediction mapping in Wadi El-Laqeita and surroundings, Central Eastern Desert, Egypt.* Arabian Journal of Geosciences, 2021. **14**(4): p. 323.
2. Ahmad, A., et al., *Flood risk modelling by the synergistic approach of machine learning and best-worst method in Indus Kohistan, Western Himalaya.* Geomatics, Natural Hazards and Risk, 2025. **16**(1): p. 2469766.
3. Akhyar, A., et al., *Deep artificial intelligence applications for natural disaster management systems: A methodological review.* Ecological Indicators, 2024. **163**: p. 112067.
4. Efraimidou, E. and M. Spiliotis, *A gis-based flood risk assessment using the decision-making trial and evaluation laboratory approach at a regional scale.* Environmental Processes, 2024. **11**(1): p. 9.
5. Alshammary, M.J., I.R. Karim, and M.Y. Fattah, *Monitoring Flood Waves Due to Overtopping: Case Study of Mosul Dam from Iraq.* Journal of Engineering, Science and Technological Trends, 2025. **2**(1).
6. Brunner, M.I., et al., *Challenges in modeling and predicting floods and droughts: A review.* Wiley Interdisciplinary Reviews: Water, 2021. **8**(3): p. e1520.
7. Islam, A.R.M.T., et al., *Predicting flood risks using advanced machine learning algorithms with a focus on Bangladesh: influencing factors, gaps and future challenges.* Earth Science Informatics, 2025. **18**(3): p. 300.
8. Afzal, J., et al., *Effects of dam on temperature, humidity and precipitation of surrounding area: a case study of Gomal Zam Dam in Pakistan.* Environmental Science and Pollution Research, 2023. **30**(6): p. 14592-14603.
9. Jain, H., et al., *Leveraging machine learning algorithms for improved disaster preparedness and response through accurate weather pattern and natural disaster prediction.* Frontiers in Environmental Science, 2023. **11**: p. 1194918.
10. Khalaf, M., et al., *IoT-enabled flood severity prediction via ensemble machine learning models.* IEEE Access, 2020. **8**: p. 70375-70386.
11. Khan, I., et al., *Climate change impact assessment, flood management, and mitigation strategies in Pakistan for sustainable future.* Environmental Science and Pollution Research, 2021. **28**(23): p. 29720-29731.
12. Khan, M., et al., *Developing a machine learning-based flood risk prediction model for the Indus Basin in Pakistan.* Water Practice & Technology, 2024. **19**(6): p. 2213-2225.
13. Kumar, V., et al., *Machine learning applications in flood forecasting and predictions, challenges, and way-out in the perspective of changing environment.* AIMS Environmental Science, 2025. **12**(1): p. 72-105.
14. Shaikh, T.A., et al., *A fundamental overview of ensemble deep learning models and applications: systematic literature and state of the art.* Annals of Operations Research, 2024: p. 1-77.
15. Lawal, Z.K., H. Yassin, and R.Y. Zakari. *Flood prediction using machine learning models: a case study of Kebbi state Nigeria*. in *2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*. 2021. IEEE.
16. Mahmood, S., A. Sajjad, and A.-u. Rahman, *Cause and damage analysis of 2010 flood disaster in district Muzaffar Garh, Pakistan.* Natural hazards, 2021. **107**(2): p. 1681-1692.
17. Rahman, Z.U., et al., *GIS-based flood susceptibility mapping using bivariate statistical model in Swat River Basin, Eastern Hindukush region, Pakistan.* Frontiers in Environmental Science, 2023. **11**: p. 1178540.
18. Ruichen, M., et al., *Vegetation variation regulates soil moisture sensitivity to climate change on the Loess Plateau.* Journal of Hydrology, 2023. **617**: p. 128763.
19. Shehzadi, M., et al., *Enhancing flood resilience: Streamflow forecasting and inundation modeling in pakistan.* Engineering Proceedings, 2023. **56**(1): p. 315.
20. Siddiqui, M.A., et al., *Advance Ensemble Flood Warning System: A Case Study for Nullah Lai.* Environmental Sciences Proceedings, 2023. **25**(1): p. 96.
21. Tehrany, M.S., B. Pradhan, and M.N. Jebur, *Flood susceptibility mapping using a novel ensemble weights-of-evidence and support vector machine models in GIS.* Journal of hydrology, 2014. **512**: p. 332-343.
22. Ahmed, R., *Essays on Infrastructure, Firm Productivity, Natural Disaster and Life Course Transition in South Asia*. 2017.
23. Nishtar, Z. and J. Afzal, *A review of real-time monitoring of hybrid energy systems by using artificial intelligence and IoT.* Pakistan Journal of Engineering and Technology, 2023. **6**(3): p. 8-15.
24. Waleed, M. and M. Sajjad, *Advancing flood susceptibility prediction: A comparative assessment and scalability analysis of machine learning algorithms via artificial intelligence in high-risk regions of Pakistan.* Journal of Flood Risk Management, 2025. **18**(1): p. e13047.
25. Afzal, J., et al., *A complex wireless sensors model (CWSM) for real time monitoring of dam temperature.* Heliyon, 2023. **9**(2).
26. Rajab, M.A., F.A. Abdullatif, and T. Sutikno, *Classification of grapevine leaves images using VGG-16 and VGG-19 deep learning nets.* TELKOMNIKA (Telecommunication Computing Electronics and Control), 2024. **22**(2): p. 445-453.
27. Khan, M., et al., *Assessing Agrometeorological Damage after the 2022 Pakistan Floods: Insights from Multi-Sensor Satellite Data.* Earth Systems and Environment, 2025: p. 1-21.
28. Şen, Z., *Flood modeling, prediction and mitigation*. 2018: Springer.
29. Goldberg, M.D., et al., *Mapping, monitoring, and prediction of floods due to ice jam and snowmelt with operational weather satellites.* Remote Sensing, 2020. **12**(11): p. 1865.
30. Benn, D.I., et al., *Response of debris-covered glaciers in the Mount Everest region to recent warming, and implications for outburst flood hazards.* Earth-Science Reviews, 2012. **114**(1-2): p. 156-174.