

Advancing flood susceptibility prediction: A comparative assessment and scalability analysis of machine learning algorithms via artificial intelligence in high-risk regions of Pakistan

Mirza Waleed¹  | Muhammad Sajjad^{1,2} 

¹Department of Geography, Hong Kong Baptist University, Kowloon Tong, Hong Kong SAR

²The Centre for Geo-computation Studies, Hong Kong Baptist University, Kowloon Tong, Hong Kong SAR

Correspondence

Muhammad Sajjad, Department of Geography, Hong Kong Baptist University, Office AAB1222, 15 Baptist University Road, Kowloon Tong, Hong Kong SAR.

Email: mah.sajjad@hotmail.com

Funding information

HKBU Research Grants Committee

Abstract

Flood susceptibility mapping (FSM) is crucial for effective flood risk management, particularly in flood-prone regions like Pakistan. This study addresses the need for accurate and scalable FSM by systematically evaluating the performance of 14 machine learning (ML) models in high-risk areas of Pakistan. The novelty lies in the comprehensive comparison of these models and the use of explainable artificial intelligence (XAI) techniques. We employed XAI to identify significant conditioning factors for flood susceptibility at both the model training and prediction stages. The models were assessed for both accuracy and scalability, with specific focus on computational efficiency. Our findings indicate that LGBM and XGBoost are the top performers in terms of accuracy, with XGBoost also excelling in scalability, achieving a prediction time of ~18 s compared to LGBM's 22 s and random forest's 31 s. The evaluation framework presented is applicable to other flood-prone regions and highlights that LGBM is superior for accuracy-focused applications, while XGBoost is optimal for scenarios with computational constraints. The findings of this study can assist in accurate FSM in different regions and can also assist in scaling up the analysis to a larger geographical region which could assist in better decision-making and informed policy production for flood risk management.

KEYWORDS

flood risk, flood susceptibility mapping, floods, machine learning, remote sensing

1 | INTRODUCTION

Globally, 1.81 billion people (23% of the global population) are exposed to high flood risk, which is expected to increase to 2.3 billion in 2050 (Rentschler et al., 2022).

While floods are inevitable and affect millions of people every year, developed countries overcome floods with good management, strong infrastructure, and data-driven policies. Conversely, developing countries suffer more as they are ill-equipped, have limited resources, inadequate

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2024 The Author(s). *Journal of Flood Risk Management* published by Chartered Institution of Water and Environmental Management and John Wiley & Sons Ltd.

infrastructure, and higher population density in flood-prone areas (Devitt et al., 2023). Economically, alone in 2020, floods caused nearly 105 USD billion in damages, which is projected to rise to 150 USD billion in 2050 (World Meteorological Organization, 2022). Pakistan, which globally ranks 5th in population and 8th in terms of climate vulnerability, is a country that has faced cataclysmic floods in recent years (Eckstein et al., 2021). The unique geography of Pakistan makes it susceptible to heavy monsoon rains and riverine flooding (Sajjad et al., 2023). For instance, due to the recent 2022 floods in Pakistan, the economic loss exceeded 15.2 USD billion, affecting 33 million people among which 1730 died (World Bank, 2022).

Flood susceptibility (FS) depicts the inherent vulnerability of an area to flooding, indicating the likelihood of it being affected by a flood event. In the face of increasing global challenges posed by floods, flood susceptibility mapping (FSM) has emerged as a crucial component of effective flood risk management and disaster preparedness, providing critical insights for policymakers, urban planners, and stakeholders, enabling them to make informed decisions and implement targeted interventions (Seleem et al., 2022). By effectively identifying high-risk areas and comprehending the contributing factors to flood vulnerability, such as geography, climate, and land use, proactive measures can be taken to mitigate the impacts of flooding, safeguard communities, and minimise economic losses (Towfiqul et al., 2021; Zhao et al., 2019). Integrating FS analysis into planning processes empowers societies to enhance resilience and promote sustainable development in flood-prone regions (Serdar et al., 2022).

Conventional FSM methods involve integrating various geospatial data, hydrological modelling, and advanced analytical techniques. Using multiple flood influencing factors, flood risk maps are prepared, which assist in identifying regions where significant impacts are more likely (Schumann et al., 2009). Furthermore, hydrological models are employed for future simulation, which simulates flood risk scenarios based on historical data and future climate projections (Tsakiris, 2014). Recent advancements in machine learning (ML) have significantly enhanced FSM. However, many existing studies focus on limited geographic areas or specific models, lacking a comprehensive comparative analysis. Studies such as Karakas et al. (2023) and Luu et al. (2021) have demonstrated the potential of ML in FSM but often overlook scalability and broader applicability. Moreover, the validation of these models against real-world flood events remains inadequate. This study addresses these shortcomings by evaluating the performance and scalability of 14 different ML models in FSM across diverse and high-risk regions of Pakistan.

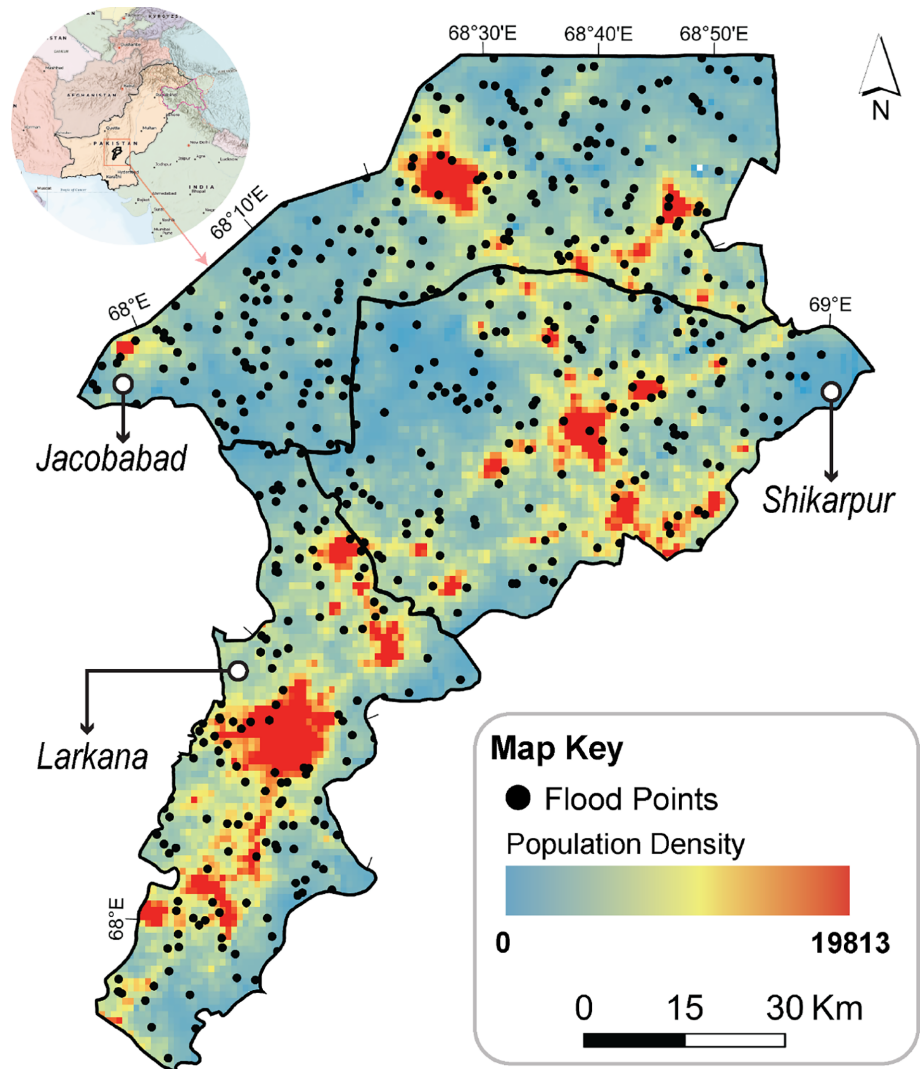
Flood-influencing factors are generally categorised into meteorological (i.e., the frequency, duration, and intensity of precipitation events), hydrological (i.e., the characteristics of topographic surface, elevation, river, basin, etc.), and anthropogenic factors, such as land use, urbanisation, and vegetation change (Mudashiru et al., 2021; Nkwunonwo et al., 2020; Zhang et al., 2023). Thus, relying on ML-based modelling of FSM through geospatial technology, data integration, and advanced modelling techniques provides a comprehensive understanding of flood-prone areas, enabling informed decision-making (Mudashiru et al., 2021; Nkwunonwo et al., 2020; Notti et al., 2018; Schumann et al., 2009; Seleem et al., 2022; Serdar et al., 2022; Teng et al., 2017; Tsakiris, 2014).

The growing global flood risk, with 1.81 billion people currently exposed and projections suggesting an increase to 2.3 billion by 2050, underscores the urgent need for effective FSM. Existing FSM methods often fail to scale effectively or consider local variations in geography and climate, leading to suboptimal results. Our study aims to bridge these gaps by comparing the performance of multiple ML models in FSM, particularly focusing on their scalability and applicability in high-risk regions of Pakistan. Scaling the modelling to a larger geographical region requires extensive knowledge about model working, training time, prediction time, and computational requirements, which are currently not documented well. Our study aims to address the critical gaps in current FSM approaches by providing a comprehensive comparative analysis of 14 ML models. We focus on evaluating their scalability, accuracy, and applicability in high-risk regions, specifically in Pakistan. This research not only benchmarks the performance of these models but also offers insights into their suitability for large-scale and diverse geographic applications, ultimately contributing to more effective flood risk management.

2 | SITE DESCRIPTION

Among developing countries, Pakistan has faced deadly floods in recent years due to climate change, geography, poverty, and negligence from the concerned authorities (Sajjad et al., 2023; Ullah et al., 2022). A recent study by Akhtar et al. (2023) found that for the recent 2022 flood in Pakistan, Shikarpur, Jacobabad, and Larkana were the three high-priority regions due to their significant population exposure and extensive flood extent. Similarly, Pakistan's previous flood impacts have been devastating in terms of social, economic, and environmental damages (i.e., the flood of 2010 and 2014). So as a case study, we focused on these critical regions to provide further insight

FIGURE 1 Study area map of three districts (Larkana, Shikarpur, and Jacobabad) with population density. The population density data is taken from www.worldpop.org.



into flood susceptible areas and, at the same time, compared various models' efficiency in terms of accuracy and scalability.

The three key regions (i.e., Shikarpur, Jacobabad, and Larkana) are situated in the South of Pakistan, specifically in Sindh province (Figure 1). The study area covers an area of $\sim 7164 \text{ Km}^2$, with Jacobabad, Larkana, and Shikarpur each having 2691 Km^2 , 1921 Km^2 , and 2551 Km^2 of area, respectively. The total population of the study area is 877,021, with which Larkana has the highest population of 490,508, whereas Jacobabad and Shikarpur have 195,437 and 191,076, respectively.¹ The study area is in a subtropical region with an elevation profile between 38 and 166 m above sea level. The province of Sindh lies between two monsoons, namely the southwest monsoon (from the Indian Ocean) and the northeast, also known as the retreating monsoon. While the average annual rainfall of Sindh province is only 230 mm, the region has suffered dramatically due to historical flood events resulting in catastrophic damage (Atif et al., 2021).

3 | DATA AND METHODS

3.1 | Data acquisition

The overall methodology of this study is divided into six sections, which include data acquisition, feature preparation, training and validation samples, ML modelling, accuracy assessment, and flood risk assessment (Figure 2). The study starts with using data from multiple sources, including Landsat-8 (United States Geological Survey²), digital elevation model (Hawker et al., 2022), vector data (i.e., shapefile data containing Pakistan administrative boundaries, river, basin, and roads shapefiles³), flood inventory data, and Climate Hazards Group InfraRed Precipitation with Station data (CHIRPS) rainfall data (Wahyuni et al., 2021)—see Table 1. These datasets are then used to prepare 11 flood influencing factors, which are preprocessed and supplied to ML models. Lastly, various accuracy assessment metrics are used to evaluate the performance of each ML model in FSM.

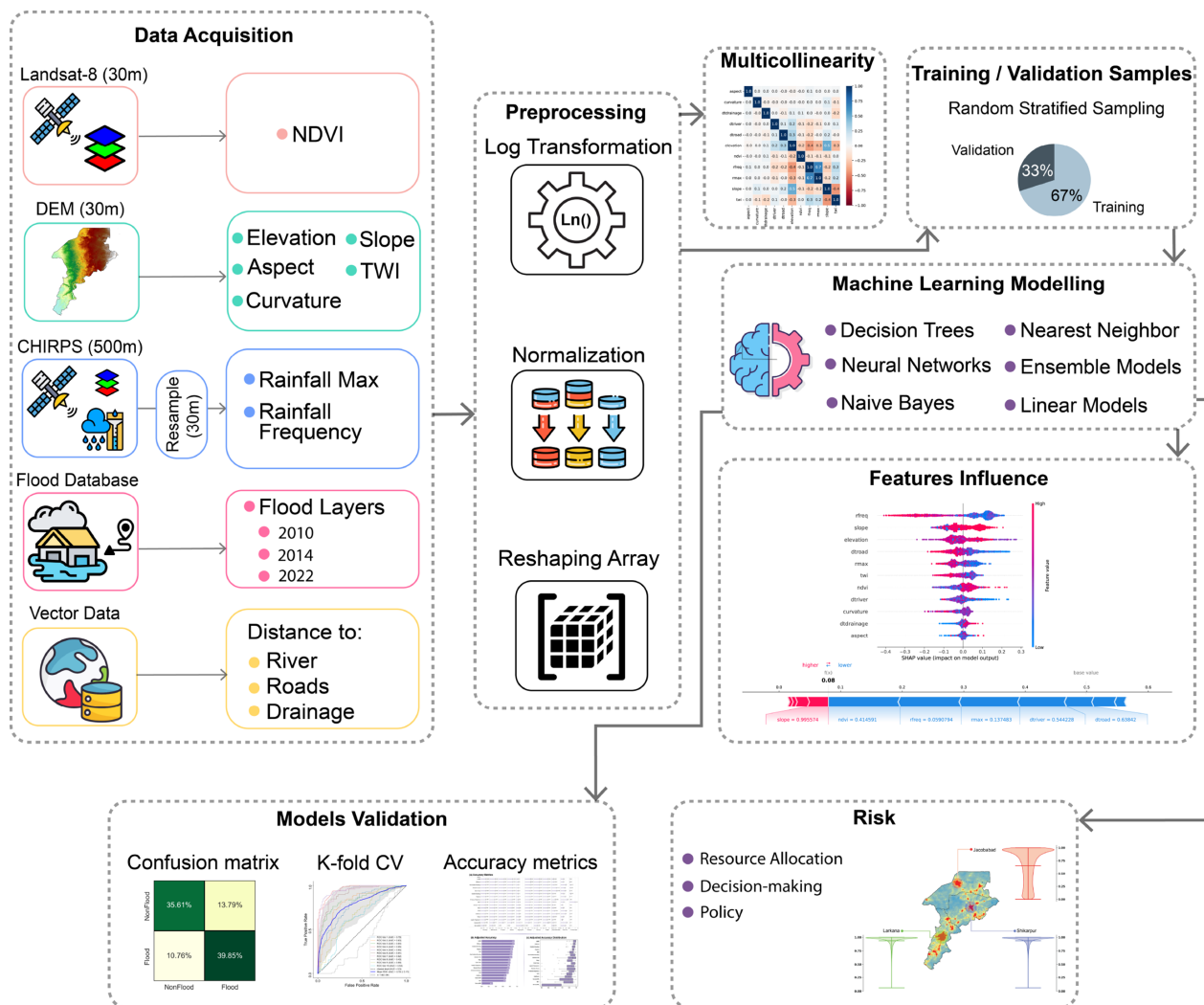


FIGURE 2 Methodology flowchart adopted in this study.

3.2 | Features preparation

3.2.1 | Flood inventory preparation

A flood inventory database contains detailed information on the spatial and temporal extent of historical flood events (Adhikari et al., 2010). Due to the unavailability of a comprehensive flood database in Pakistan, we created our inventory using flood layers for 2010, 2014, and 2022 flood events provided by existing studies and institutions (see Table 1). By combining these three layers, a new flood inventory layer was formed, which depicts the characteristics of historical flood events in the study area. Next, for creating training and validation samples for modelling, the intersection of layers was performed, and the final image was taken which filters only those pixels common in all three flood events. For sampling, a total of 4000 samples were generated using an equally random stratified technique (Parsons, 2017) ensuring an equal

proportion (i.e., 2000 samples for flood and 2000 for non-flood locations), which benefits binary classification models in understanding each class without bias (Kaya & Gündüz Ögüdücü, 2018). Then, the samples were divided into three-to-one ratios (i.e., 67% of samples were used further for training the models, whereas the remaining 33% were used for testing).

3.2.2 | Flood conditioning factors

While recent studies have used up to 20 flood-influencing factors (Seydi et al., 2023), they often induce multi-collinearity and thus induce bias in the model. Therefore, based on the least correlated variables from the literature (Pandey et al., 2021; Seleem et al., 2022; Towfiqul et al., 2021), we used 11 flood-influencing factors shown in Figure 3. Elevation, which is a main conditioning factor for floods, is globally available in different end

TABLE 1 Datasets used in this study and their source information.

Dataset name	Resolution (year)/type	Justification
Drainage	Shapefile	Lehner and Grill (2013)
Elevation	30 m (2022)	Hawker et al. (2022)
Flood layers (2010 and 2014)	Shapefile	www.unitar.org ^a
Flood layer 2022	Shapefile	Akhtar et al. (2023)
Land-use	10 m (2020)	Karra et al. (2021)
CHIRPS rainfall	5.5 km (2010–2022)	Funk et al. (2015)
River boundary	Shapefile	Lehner and Grill (2013)

^aFlood Layers (2010 and 2014) Shapefile data provided by UNITAR upon personal request. Researchers can visit www.unitar.org and email them to request access to the data.

products, each with a 30–90 m spatial resolution. Four variables were derived from elevation, including aspect, slope, curvature, and topographic wetness index (TWI) (Towfiqul et al., 2021). The aspect shows the direction of the slope, with values ranging between 0 and 360 (clockwise). The aspect values 0, 90, 190, and 270 represent north, east, south, and west, respectively. The slope is the measure of the steepness of a ground surface and is calculated from DEM in either degree or percentage units. Curvature, (also known as profile curvature) is a quantity parallel to the direction of the maximum slope and indicates the acceleration and deceleration of flow through the surface. TWI is an index quantity that highlights the terrain-induced variations in soil moisture properties.

For meteorological factors, we used daily precipitation between 2010 and 2022 from CHIRPS and prepared two variables, namely frequency of rainfall (greater than 10 mm), and maximum precipitation. For frequency of precipitation, Google Earth Engine (GEE) was used, and a total of 49 events were recorded in the study area. The frequency of the precipitation dataset was then downsampled to 30 m spatial resolution using a 4 by 4 majority filtering resampling technique.⁴ The maximum precipitation image was also first processed in GEE and was then exported into local ArcGIS Pro⁵ workflow, where it was downsampled to 30 m using the Bayesian kriging regression prediction (EBKRP) technique, which involves using elevation data to downscale precipitation (Ali et al., 2021). For anthropogenic factors, the normalised difference vegetation index (NDVI) was prepared using the Landsat-9 dataset for the year 2022, whereas the distance to major roads raster was prepared using Euclidian distance (Waleed, Sajjad,

Acheampong, & Alam, 2023). Besides these, the distance to rivers and drainage was also prepared using Euclidian distance, at 30 m spatial resolution.

3.2.3 | Data preparation and preprocessing for flood variables

Data conversion and overview of machine learning models

For ML modelling using remotely sensed data, the conventional data is reshaped into a numerical 2D array, organising it to represent spatial and spectral dimensions (Fang et al., 2021). Before that, it is important to transform and normalise the data to make it utilisation-ready in ML models (Ali et al., 2021; Ha et al., 2021). Details on these steps are provided in the following sections.

Log transformation of the data

Log transformation is the widely used data processing technique in ML and is defined as the process of transforming non-normally distributed data using either a natural logarithm or a logarithm with a specific base to the data (Ha et al., 2021). However, since the logarithm can be applied to only positive values and specific data, such as NDVI, can contain negative values, a scaling factor is often added to the equation to make values positive (Seleem et al., 2022). The overall equation used to perform log transformation is provided as Equation (1).

$$y_0 = \partial \times \log(x + \delta), \quad (1)$$

where y_0 represents the transformed value, ∂ represents the scaling factor applied, x represents the original value, and lastly, δ represents the constant value added.

Feature normalisation

Normalisation is a crucial data processing technique in ML modelling that adjusts values to a standard scale. This step is essential because certain models are sensitive to the scale of input features, yielding better results when data is normalised. There are two common approaches to normalisation based on data distribution: positive normalisation and negative normalisation.

Positive normalisation scales data to a specific range, typically between 0 and 1, which is useful when the absolute values or relative magnitudes of features are important. Conversely, negative normalisation, also known as z-score normalisation, transforms data to have a mean of 0 and a standard deviation of 1. This method is advantageous when the distribution shape or presence of outliers are significant factors, as it centres the data and equalises the scales of different features. The equations for positive

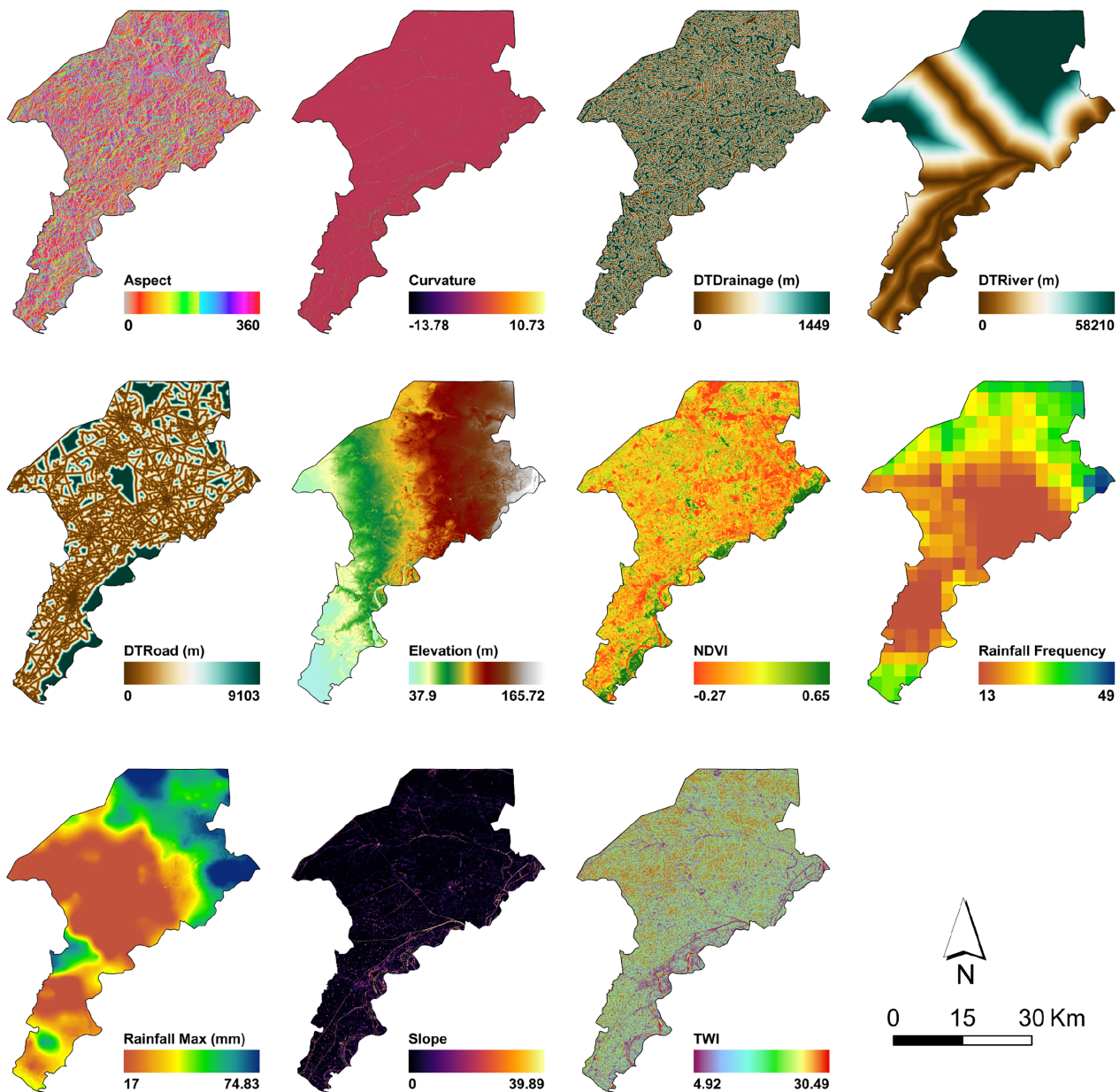


FIGURE 3 Flood influencing factors used for flood susceptibility mapping (FSM).

and negative normalisation are provided as Equations (2) and (3), respectively.

$$\gamma_1 = \left[\frac{x - \min(x)}{\max(x) - \min(x)} \right], \quad (2)$$

where γ_1 represents the positive normalised value, x is the original value in the data, $\min(x)$ is the minimum value in the data, and lastly, $\max(x)$ is the maximum value in the data.

$$\gamma_2 = \left[\frac{\max(x) - x}{\max(x) - \min(x)} \right], \quad (3)$$

where γ_2 represents the normalisation of negatively contributing conditioning factors, x is the original value, $\max(x)$, and $\min(x)$ are the same as Equation (2).

While normalisation transforms values on a standard scale, it does not address data skewness or the influence of extreme values effectively. Log transformation helps in reducing the influence of outliers and makes the data more symmetrical, thereby improving the model's performance and stability (Ha et al., 2021). Normalisation alone is often insufficient because it can leave highly variable data skewed, impacting the accuracy and interpretability of ML models. By combining log transformation with normalisation, we achieve a more robust data

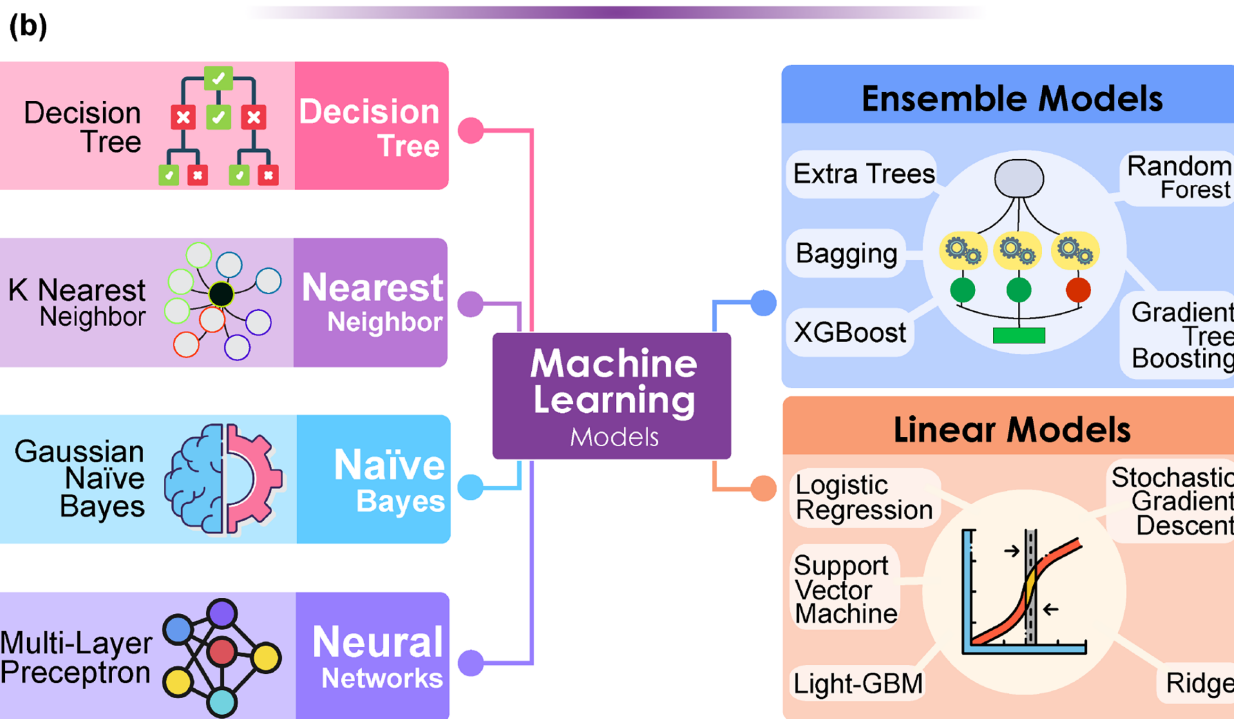
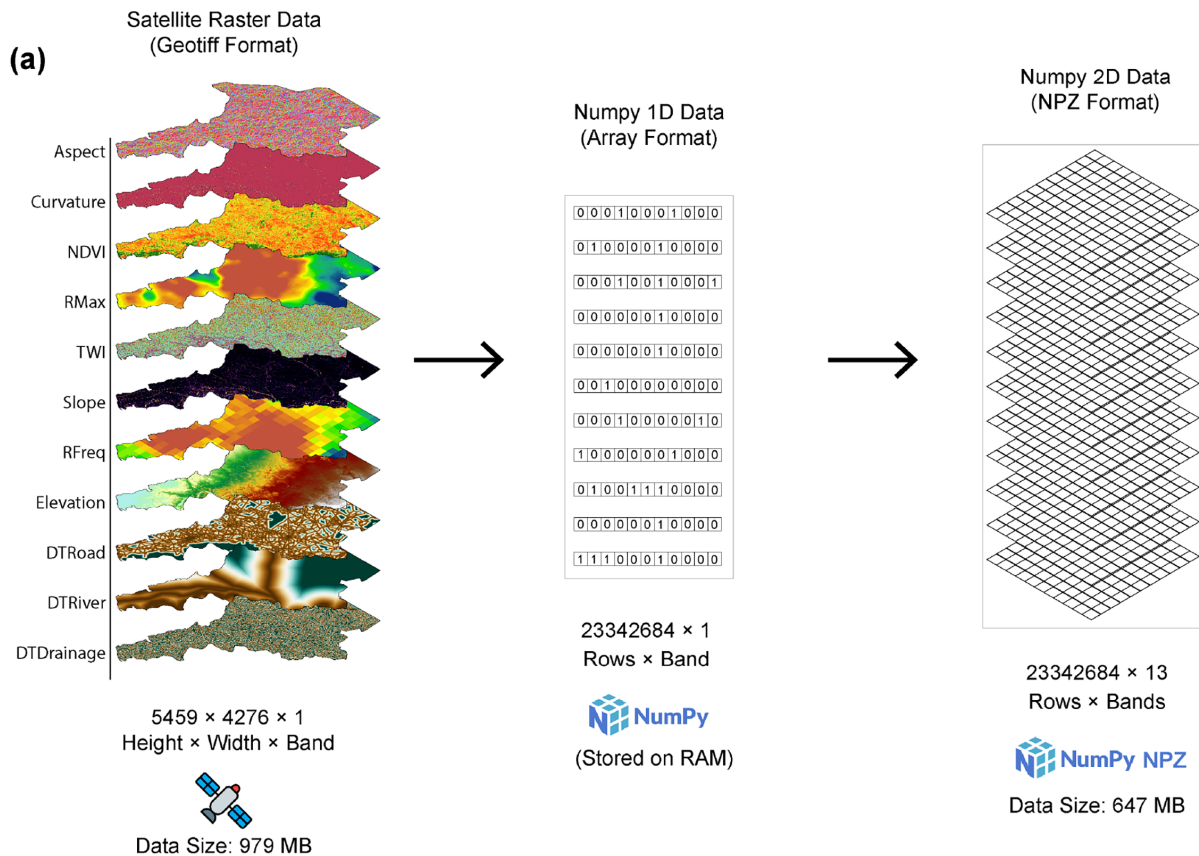


FIGURE 4 (a) Overview of features data conversion, and (b) categorisation of different machine learning (ML) models utilised in this study.

preprocessing pipeline that enhances the overall model effectiveness (Ali et al., 2021).

Figure 4a shows the flow of data conversion, where 11 features in “Geotiff” format are first converted into 1D

array format using the NumPy python package.⁶ Then, the NumPy array is reshaped into 2D, where rows depict values and columns represent bands (features). This reshaped NumPy array is then saved locally using

NumPy native compressed format. The benefit of saving in NumPy-based compressed format includes a reduction in file size (979 MB to 647 Mb), faster loading and processing speed, and minimum chances of processing errors (i.e., due to Python package conflicts). The resulting 2D structure captures spatial distribution and facilitates efficient analysis by assigning pixel values to their corresponding array positions. Furthermore, layers are stacked for multiple features to create one 2D data cube, which can be used further for model training and prediction.

Since FSM is a classification probability problem, we focused on all supervised ML models. First, six different categories of ML classification models were prepared based in Scikit-Learn,⁷ namely decision trees, nearest neighbour, Naïve bayes, neural networks, ensemble models, and linear models. For each category, we identified all available base ML models, which could generate classification probabilities. As a result, 14 different ML models were identified and further used for FSM. The overall model overview and individual model names are shown in Figure 4b.

3.3 | ML model validation

Conventionally, ML model validation is performed using multiple accuracy metrics that depict the model's performance regarding true and false identified labels. To conduct a comprehensive validation, we employed all available accuracy metrics, including overall accuracy, precision, recall, F1 score, Jaccard score, net log loss, receiver operating characteristic (ROC) curve, and area under the curve (AUC). This extensive validation process assessed the reliability and predictive capabilities of the ML models, ensuring the accuracy of FSM. The details of each metric are provided in Table 2.

4 | RESULTS

4.1 | Models validation

The results of the accuracy assessment of ML models are presented in three parts. In the first part, the distribution of accuracy assessment metrics is provided (Figure 5). Among all models, LGBM and XGBoost take first and second places, respectively. Among these two models, LGBM performs better in terms of accuracy (~ 0.84), whereas XGBoost performs better in terms of scalability measured by prediction time (18 s). Recent studies propose an adjusted accuracy method for assessing the effectiveness of models (Ali et al., 2021). In adjusted accuracy, all possible accuracy metrics for the model are evaluated,

and then their mean value is taken. The resulting adjusted accuracy based on the previous 11 accuracy metrics is shown in Figure 5b, which ranks LGBM and XGBoost at the top first position, while Gradient Boosting and Bagging at second and third position, respectively. Notably, the popular model random forest (RF), widely used in previous studies, showed satisfactory accuracy ($F1s = 0.83$, adjusted accuracy = 0.79) but showed poor scalability with a prediction time of up to 31 s.

Lastly, when evaluating the mean adjusted accuracy, it is crucial to document the variation in the range of accuracy metrics values and highlight potential outliers, as presented in Figure 5c. It is evident that the density of most of the models, including LGBM and XGBoost, is shifted towards the right indicating overall better performance. Moreover, except for decision tree, and GaussianNB models, all models showed sustained performance with a minimum value greater than 0.5.

To document correctly classified and misclassified samples, a confusion matrix (Figure 6) is used to observe the underfitting and overfitting of the ML model (Waleed & Sajjad, 2022). Among all models, GaussianNB, LGBM, Gradient Boosting, and Decision Tree attained the highest true positive (TP) values of 45.9%, 43.79%, 43.33%, and 43.33%, respectively, indicating a good agreement between actual and predicted flood points. However, decision tree and GaussianNB showed high false positive (FP) values, indicating overfitting of models resulting in overestimation. Lastly, minimum FP and false negative (FN) values are observed in the XGBoost, LGBM, and RF models, indicating their superiority over other models in handling overfitting, overestimation, and misclassification.

The ROC and AUC are widely used to assess the effectiveness of binary ML models and for in-depth comparison of various models. The ROC summarises the trade-off between the TP rate and FP rate, considering multiple classification thresholds, also known as the fold or K-fold validation process. It also provides the perfect balance between precision and recall, emphasising the overall predicting power of a model given the variation in training data. The AUC value is used to compare the models, which is the mean value of all the ROC epochs. The AUC-based model comparison is a standardised approach, and generally, the model with higher AUC can better discriminate and predict the label class (Seleem et al., 2022). Among all the models, LGBM, XGBoost, RF, and Gradient Boosting performed better (Figure 7). Each model demonstrated a mean AUC value of 0.84, with the highest AUC recorded at 0.99 and the lowest at 0.59.

The similarity in the mean AUC values across these four models suggests that all of them exhibit comparable performance levels when prioritising accuracy.

TABLE 2 Details of accuracy assessment metrics used in this study.

Name	Description	Equation	Justification
Accuracy	Also known as overall accuracy measures the proportion of correctly classified instances in relation to the total number of instances. It provides a high-level assessment of the model's predictive accuracy and its ability to classify areas correctly	$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$	Olofsson et al. (2014)
F1-score	Based on precision and recall, it evaluates both false positives and false negatives and thus provides a balanced evaluation of model performance	$F1s = \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$	Waleed and Sajjad (2022)
Jaccard score	It calculates the ratio of the intersection between the predicted and actual flood-prone areas to their union	$Js = \frac{TP}{TP+FN}$	Maxwell et al. (2021)
Net log loss	Net log loss is a probabilistic measure that assesses the model's ability to estimate the likelihood of flood occurrence. It calculates the logarithmic loss of predicted probabilities compared to the true probabilities, penalising inaccurate predictions. A lower net log loss indicates better predictive performance	$NLL = -\frac{1}{N} \times \sum [y \times \log(p) + (1-y) \times \log(1-p)]$	Gao et al. (2022)
Precision	Precision is the ratio of correctly classified flood-prone areas to the total number of areas classified as flood-prone. It indicates the model's ability to minimise false positives, ensuring that areas classified as flood-prone are indeed at risk	$\text{Precision} = \frac{TP}{TP+FP}$	Waleed, Sajjad, Shazil, et al. (2023)
Recall	Recall, also known as sensitivity or true positive rate, represents the proportion of correctly classified flood-prone areas to the total number of actual flood-prone areas. It measures the model's ability to detect and capture all areas that are truly at risk	$\text{Recall} = \frac{TP}{TP+FN}$	Waleed, Sajjad, Shazil, et al. (2023)
ROC and AUC	ROC and AUC analysis evaluate the model's ability to discriminate between flood-prone and non-flood-prone areas across different classification thresholds. It comprehensively assesses the model's performance at various probability cutoffs	The ROC curve is constructed by plotting the true positive rate (TPR) against the false positive rate (FPR) at different classification thresholds	Seleem et al. (2022)

Abbreviations: FP, false positives (incorrectly predicted flood-prone areas); FN, false negatives (incorrectly predicted non-flood-prone areas), N , total number of instances; p , predicted probability; TP, true positives (correctly predicted flood-prone areas); TN, true negatives (correctly predicted non-flood-prone areas); y , true label (0 or 1).

Additionally, these models consistently outperformed similar models regarding classification accuracy. Notably, the highest AUC of between 0.97 and 0.99 achieved by all four models (LGBM, XGBoost, RF, and gradient boosting) indicates their exceptional discriminatory power and ability to distinguish between flood and non-flood classes effectively. Therefore, these findings highlight the competitive performance of the LGBM, XGBoost, RF, and gradient boosting models, affirming their suitability for the FSM task.

4.2 | Comparison of machine learning models

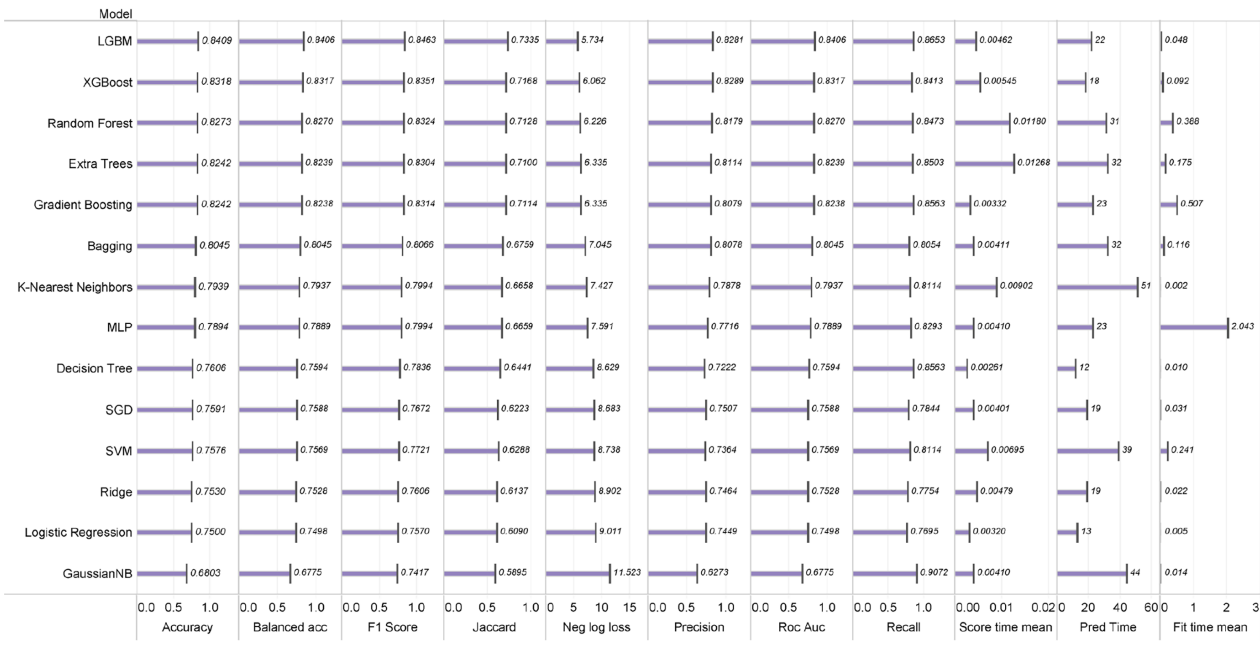
To enhance the clarity and facilitate comparative analysis of the ML models evaluated in this study, we have

summarised the advantages and disadvantages of each model based on accuracy assessment findings. This comprehensive overview helps in understanding the strengths and limitations of each model in the context of FSM and is provided as Table 3.

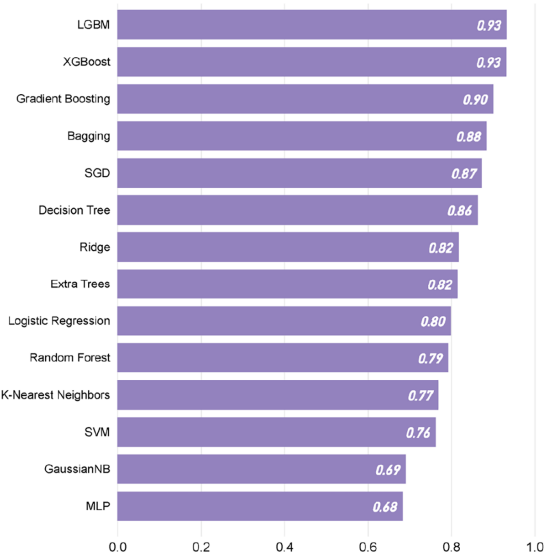
4.3 | Dominant factors controlling flood susceptibility

Correlation analysis is performed to assess the multicollinearity between features, which is given in Figure 8a. No multicollinearity is observed between variables given the upper and lower correlation threshold of -0.75 to 0.75 , respectively. Furthermore, the only highest correlation of 0.7 is observed between rainfall frequency (rfreq) and

(a)



(b)



(c)

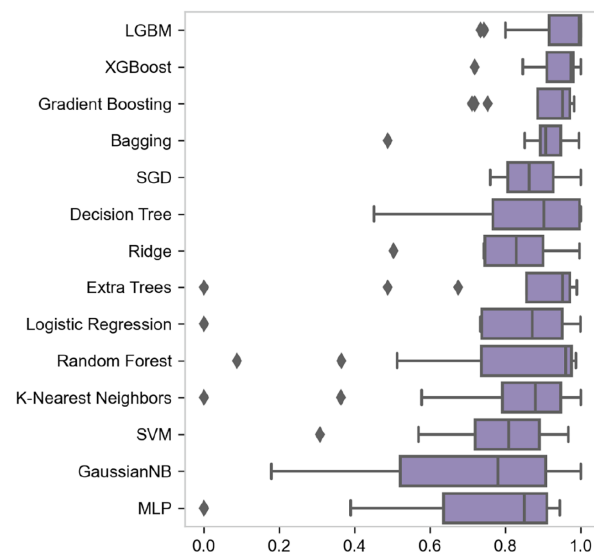


FIGURE 5 Accuracy assessment of FSM models. (a) Accuracy metrics for all models, using 11 accuracy metrics. (b) Adjusted accuracy based on an average of all accuracy metrics, and (c) Boxplot showing the distribution of adjusted accuracy. (a) A bar plot of individual accuracy metrics for each model, (b) the adjusted accuracy based on the mean of each metric, and lastly, (c) the violin plot for the distribution of adjusted accuracy for each model. (a) Is sorted based on the F1s.

maximum rainfall (rmax), which is reasonable given that both variables highly contribute to the variation in flood occurrence (Seleem et al., 2022).

Figure 8b, c shows the feature influence force plot and feature importance ranking, respectively, estimated with the SHapley Additive exPlanations (SHAP) method (Lundberg & Lee, 2017), and using the highest accurate LGBM model. The SHAP is a theoretical framework that explains various variables' role in ML training and

output prediction by taking the average marginal contribution of each feature associated with all available coalitions. Figure 8b shows the feature influence force plot, in which the base value of 0.5 (50% susceptibility, in the absence of any feature) is indicated based on the average of prediction in the training dataset. The higher values (in red) highlight a direct influence of variable on the training model (i.e., slope with a value of 0.9 corresponds to red colour indicating that slope has a



FIGURE 6 Confusion matrix for each ML model. In each confusion matrix, the left horizontal rows show the actual percentage of samples, whereas the bottom vertical columns show the predicted percentage of samples. The top left section in the matrix indicates the correct classified non-flood pixels and is called true negative (TN). The top right section shows the incorrect classified non-flood into flood class and is called false positive (FP). Similarly, the bottom left, and right sections show false negative (FN) and true positive (TP) values, respectively.

higher influence in flood susceptibility). Similarly, blue values on the right show lower influence features, with the width of the bar showing the intensity at which features change the output model. Among lesser influencing features, NDVI, rfreq, rmax, dtriver, and dtroad illustrate an inverse relationship with flood probabilities (i.e., the increase in value of these features reduces the FS probability).

Lastly, Figure 8c shows a feature importance ranking plot, in which the Y-axis shows the most important features for model prediction (output) in descending order whereas the X-axis shows model impact and colour signifies the intensity of feature value. Among all features, rfreq being on the top signifies being the highest influencing feature for FS prediction. Specifically, the blue values of rfreq on the right side show that the lower (negative)

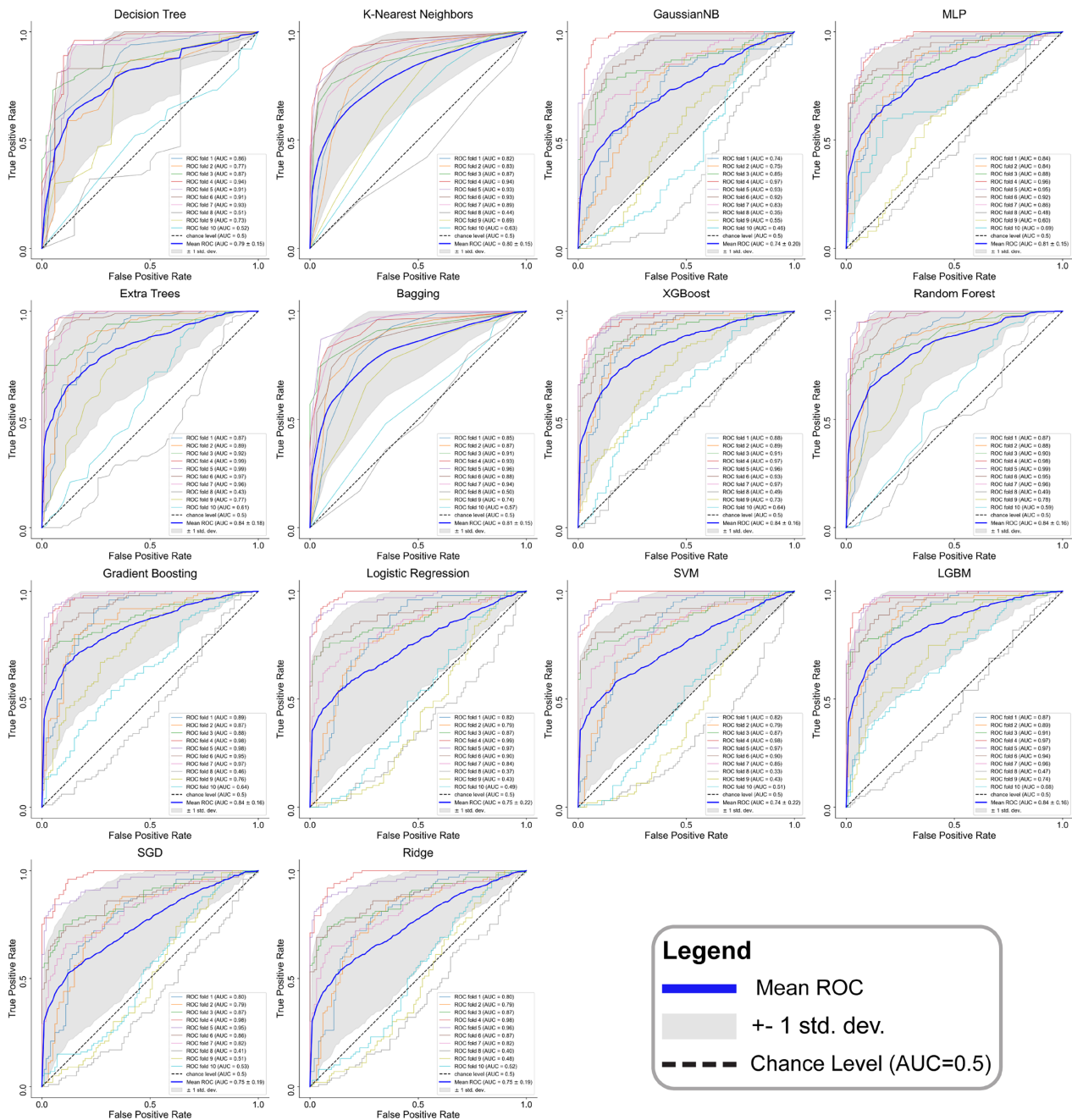


FIGURE 7 Receiver operating characteristic (ROC) curve and area under the curve (AUC) plots for each model. Here, the blue line indicates mean ROC, the grey background indicates ± 1 standard deviation, the dashed line indicates chance level with AUC of 0.5, and the coloured lines indicate ROC of 10 epochs.

values of rainfall frequency have a high impact on model output. Similar to Figure 8b, the slope feature stood second, with the majority of values as red on the right side indicating that a higher slope directly influences the model.

4.4 | Geographical disparities in flood susceptibility

Figure 9 shows FS maps for each model, in which values near 0 (light purple colour) indicate no or minimal FS

level whereas values near 1 (dark purple) signified a high FS level. In general, all FS maps follow the same spatial trends (i.e., the northern area being the high flood-susceptible region) except for Gaussian NB, which shows overestimation (also evident in Figure 6). Among the highly accurate four ML models, LGBM and XGBoost follow similar spatial patterns of having western and northern regions indicated as highly susceptible, whereas RF and Gradient tree-based maps show slight underestimation in south-east regions. Except for LGBM, other models in the Linear Models category (Figure 4) showed

TABLE 3 Summary table showing the advantages and disadvantages of ML models for FSM based on accuracy assessment.

Model	Advantages	Disadvantages
LGBM	Highest F1 score (0.846), highest accuracy (0.841), fast training and prediction (22 s)	Sensitive to overfitting, requires careful parameter tuning
XGBoost	High F1 score (0.835), high accuracy (0.832), fast prediction time (18 s)	Computationally intensive, slower with very large datasets
Random forest	High F1 score (0.832), high precision (0.818)	Requires more computational resources, high prediction time (31 s)
Gradient boosting	High F1 score (0.831), robust accuracy (0.824)	Computationally intensive, sensitive to overfitting
Extra trees	High F1 score (0.830), high recall (0.850)	High negative log loss (6.335), slower prediction time (32 s)
Bagging	High F1 score (0.807), balanced accuracy (0.804)	High negative log loss (7.045), slower prediction time (32 s)
MLP	Good F1 score (0.799), high recall (0.829)	High negative log loss (7.591), slow prediction time (23 s)
K-nearest neighbour	Good F1 score (0.799), effective recall (0.811)	High prediction time (51 s), sensitive to irrelevant features
Decision tree	Good recall (0.856), fast training	Prone to overfitting, lower accuracy (0.760)
SVM	Robust to overfitting, effective recall (0.811)	High memory and computational power requirement, high prediction time (39 s)
Ridge	Prevents overfitting, balanced accuracy (0.753)	Requires feature scaling, moderate performance
Logistic regression	Fast training and prediction (13 s), good F1 score (0.757)	Assumes linear relationships, less effective with complex data structures
SGD	Prevents overfitting, balanced accuracy (0.759)	Requires feature scaling, moderate performance
GaussianNB	Fast and efficient, high recall (0.907)	Assumes feature independence, lower accuracy (0.680)

the road, and drainage network as the least susceptible regions and the network is visible in Figure 9. Lastly, some models including decision tree, K-nearest neighbour, and bagging showed underestimation and did not follow the patterns of the highly accurate flood models. The FS results indicate that the whole region is highly susceptible to flood, with Shikarpur and Jacobabad being the highest affected. Comparatively, Jacobabad is more flood-susceptible than Shikarpur. The cities epicentre of Jacobabad, Shikarpur, and Larkana fall into high FS regions, indicating larger flood impacts.

4.5 | Deployment of the most optimal model

To access the variation in FS for each city in the context of the utilisation of the outcomes for policy and decision-making through the practice of the highest-ranking ML algorithm, we assess the state of FS for key populated regions in the study area. Violin plots are prepared based on the most accurate (LGBM) algorithm, and data inside a 7 Km² buffer around the population centre point of each city. Figure 10 shows the population density map with city buffers, along with three violin plots for Jacobabad, Larkana, and Shikarpur. It is observed that both Larkana and Shikarpur showed average FS of 0.9 indicating that most of the population falls under extreme flood susceptible region. For Jacobabad however, the average FS equals 0.65, and the distribution of FS values is between 0.85 and 1. This shows that while the average FS is far lesser than the other two cities, the distribution of values near 1 indicates that the majority of the population in this city also falls under a higher FS zone. Such targeted information is useful as well as important to devise effective risk mitigation policies in the face of intensifying floods in Pakistan in particular and in other parts of the world in general.

5 | DISCUSSION

Comparative performance evaluation of ML models for FSM is a critical topic, which ensures the effectiveness and reliability of derived FS maps. Recently, some studies (Dodangeh et al., 2020; Fang et al., 2021; Luu et al., 2021; Pandey et al., 2021; Pham et al., 2021; Seydi et al., 2023; Towfiqul et al., 2021) performed a comparative analysis of few ML models, however, most of them cannot provide an in-depth comparison. Furthermore, the authenticity and accuracy of the majority of these studies are questionable as many neglect crucial ML analysis stages (i.e., multicollinearity analysis, appropriate sampling

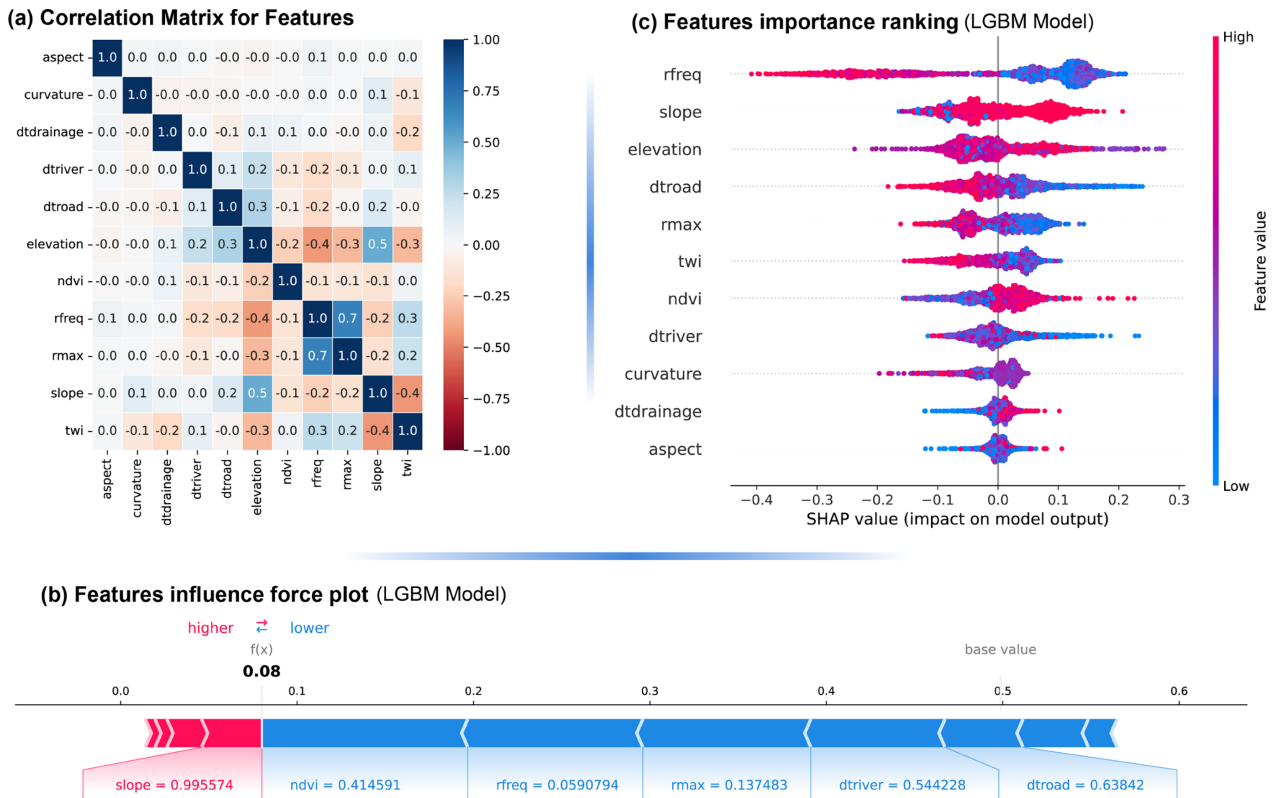


FIGURE 8 (a) Correlation matrix for all available variables, (b) feature importance in training the ML model, (c) feature influence in predicting output.

approach, comprehensive accuracy metrics for validation, choice of features, transferability assessment, scalability assessment, and lastly the selection of models) (Amiri et al., 2024; Kumar & Singh, 2024; Shah & Ai, 2024). Additionally, there has been a recent shift in studies focusing on the ML model transferability. In this context, this study systematically compared all available ML models and provides an overview of scalability performance among ML models. This provides us with a base for scaling the FS analysis to a larger geographical scale (i.e., a national or a regional scale).

The findings of this study revealed that the recently proposed models such as LGBM and XGBoost performed far better than simplified versions of base models such as decision tree, random forest, and SVM. Specifically, for accuracy assessment, both LGBM, and XGBoost attained an adjusted accuracy of 0.93, and a mean ROC of 0.84; higher than the rest of the models (Figures 5b and 7). This performance boost can be attributed to their better optimisation framework, as LGBM uses histogram-based algorithms, whereas XGBoost uses a pre-sort-based algorithm mechanism for model optimisation (Al Daoud, 2019). These mechanisms of optimisations allow these two models to be efficient in terms of faster training, reduction in memory usage, better accuracy, reduced cost for

distribution learning, support for Graphical Processing Unit learning, and lastly capability of handling large-scale data. In terms of scalability, the prediction time of LGBM (22 s) and XGBoost (18 s) is comparatively better than the most prominent models in the literature (i.e., RF (31 s) and SVM (39 s)—Figure 5a). Thus, these qualities make LGBM and XGBoost preferable for FSM, especially for scalability (i.e., extending the current FS analysis to a larger scale). In terms of reliability, Figure 6 demonstrates a confusion matrix which again confirms the minimum FP and FN values make LGBM and XGBoost superior then others in terms of handling overfitting, overestimation, and misclassification.

Previously some studies have evaluated the performance of ML models in FSM. For instance, Seydi et al. (2023) proposed an ML model based on the Cascade Forest Model. While they compared the results of this model with six other models (SVM, CART, DNN, LGBM, XGBoost, and CatBoost), they failed to justify the inter-comparison and trade-off between accuracy, transferability, and scalability. For instance, they use 17 feature variables derived from DEM, which could affect the model due to induced multicollinearity. For binary classification, a balanced dataset of labels is required, which ensures the balance in validation, whereas their study

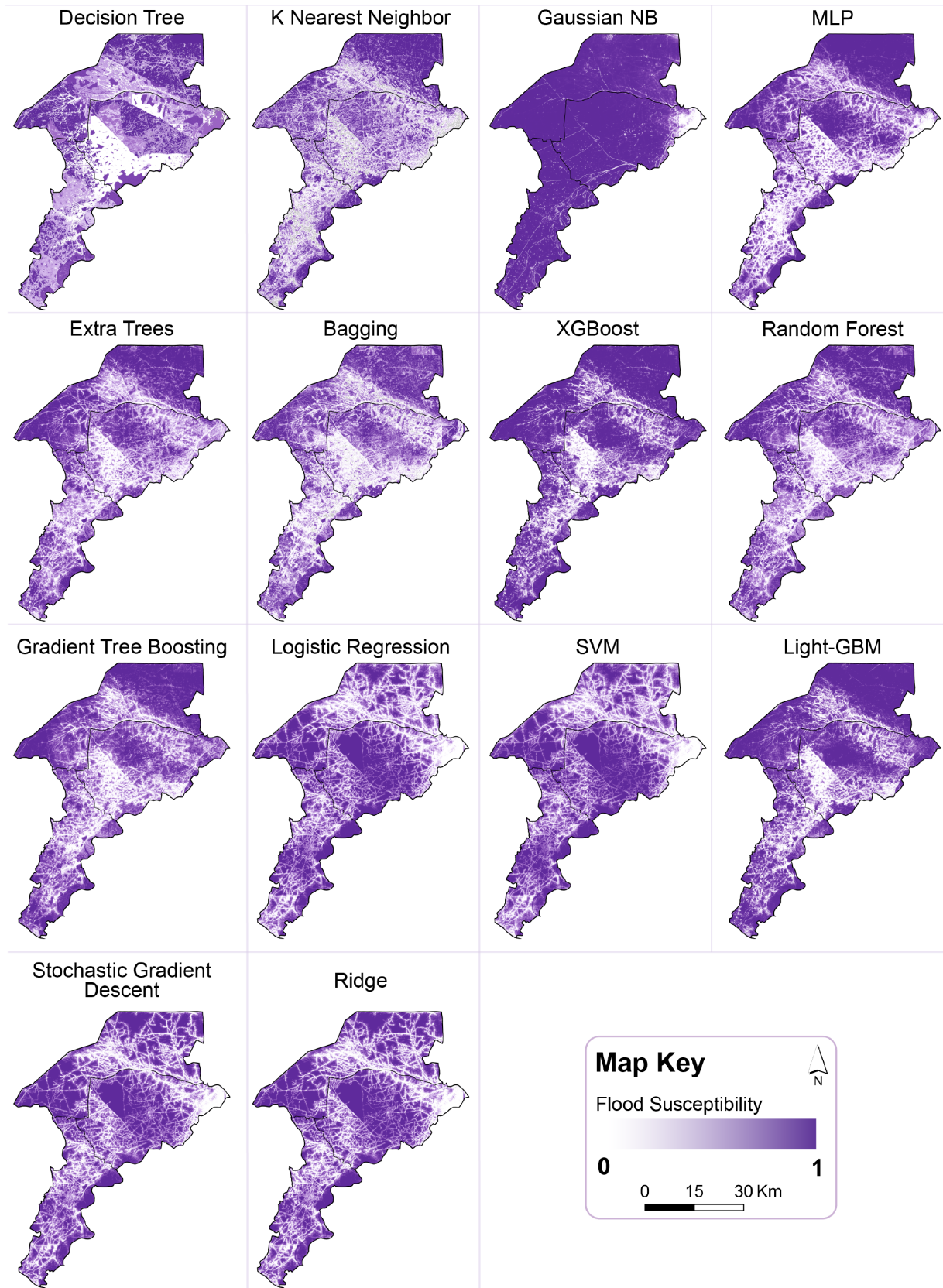


FIGURE 9 FSM maps based on the ML models assessed in this study.

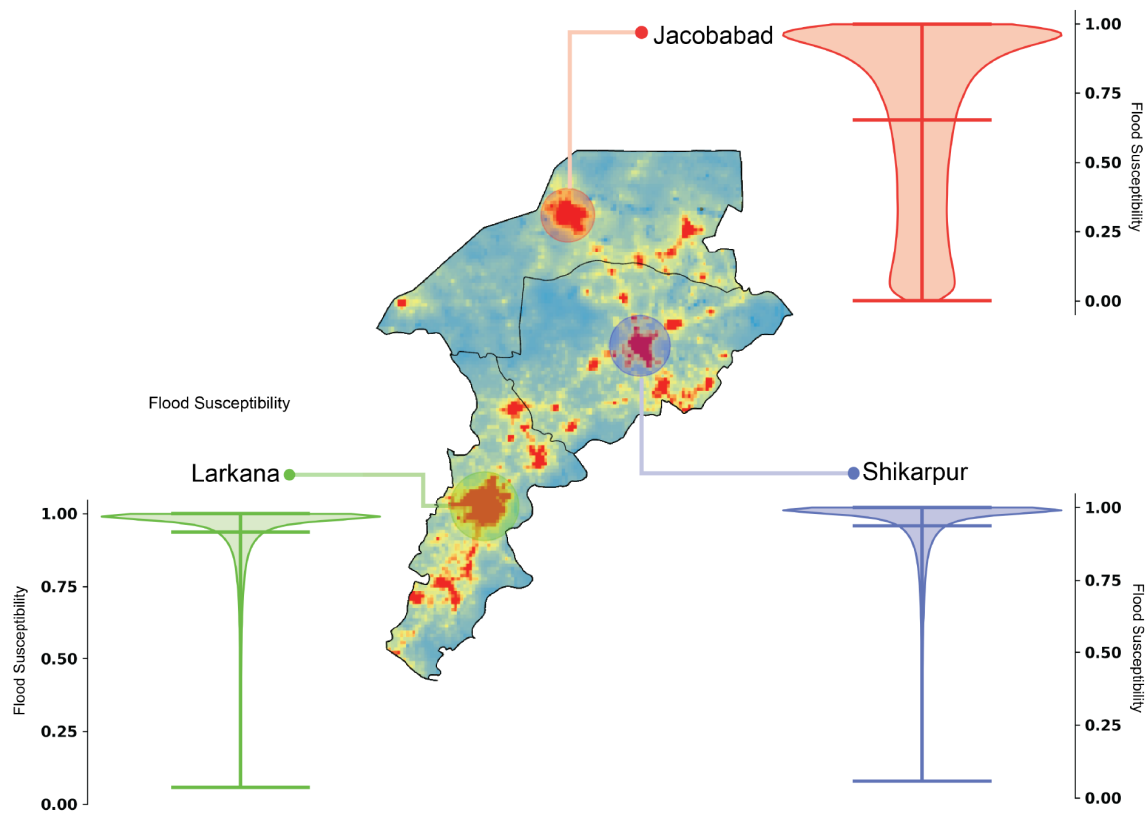


FIGURE 10 Flood susceptibility distribution in the population of each city, based on the LGBM model. The violin plots show the distribution of flood susceptibility and average flood susceptibility.

just like others, used a higher number of samples for flood class than others. For collection samples, most of the studies generate samples randomly without a sampling approach, which could create bias in model training and validation. To avoid this, we recommend using a random stratified samples approach with a custom minimum distance per sample, which not only ensures balanced sampling per class but also avoids sample clustering and thus, reduces end model bias. Furthermore, for correlation and feature importance, our study proposed employing the SHAP method, to assess the importance of each variable in both model training and model output prediction. This evaluation of feature importance at both stages is particularly helpful not only in model optimisation but also in highlighting the key features that are significant in flood risk management (Pradhan et al., 2023).

Recently, Hawker et al. (2022) recently proposed an improved version of the DEM product using a ML approach combined with the Copernicus Digital Elevation Model (CopDEM30) dataset, Light Detection and Ranging data from 12 countries, and ground-truth elevation data from 100 locations around the world for validation. They conclude that the dense vegetation-induced effect can severely affect the end product, and thus

depending on the application will have biased results. Instead, they proposed an improved version of the DEM product with such artefacts removed, which showed promising results. Therefore, in our study, we utilised Hawker et al.'s (2022) provided DEM dataset, which showed greater details at 30 m resolution as compared to previously used.

5.1 | Limitations and the way forward

While our study provided a comprehensive comparative scholarship on the state-of-the-art ML models, it is imperative to discuss limitations faced during the execution of this research. First, the lack of high-resolution datasets on a large scale makes it difficult to conduct similar comparative evaluations of ML models at different resolutions and locations. For instance, developed countries have the resources to create and host databases of high-resolution remote sensing (i.e., elevation from point cloud data (drones) and in situ datasets) (e.g., rain gauge-based rainfall measurements). This assists in providing high-resolution flood maps. Conversely, developing countries (especially Southeast Asian countries, such as Pakistan) lack such resources, hindering detailed

assessment. Hence, the only viable solution is to preprocess the available data and rely on global/regional products (i.e., elevation (30 m) and rainfall (250 km)). No doubt that doing so might induce unexpected bias in modelling due to poor data quality to a certain extent. Therefore, based on this, we believe that the overall model evaluation and comparisons can be greatly improved by incorporating high-resolution datasets especially those which can substitute currently available coarser products using the similar approach as adopted here.

Additionally, FSM requires high computational requirements due to the extensive interconnected relationship of flood influencing features, the ML model architecture, and the size of data to analyse. Hence, further advancements in terms of robustness are a potential domain to focus on. While the main objective of this study was to provide a comparative assessment of ML models in the context of efficient FSM, the comparison of all models at larger regions (national or global) can further be explored, which could provide greater insights into the response of models in different geographical and climate regions at different scales.

Based on this study, future studies can utilise LGBM or XGBoost, to scale up FSM to a national scale, which could better assist in emergency preparedness and improve the effectiveness of flood management strategies and responses by identifying high-risk areas. The integration of updated remote sensing data into our framework would provide progressive opportunities for this. Such an integration coupled with field observations and sensors' data can further assist in building real-time monitoring systems and warning mechanisms in high FS regions. Furthermore, other than the datasets used in this study, future studies can investigate the integration of socio-economic factors and community-based knowledge into FSM. For instance, the incorporation of climate projections and anthropogenic activities' simulations under the projected climate scenarios, shared socio-economic pathways, and representative concentration pathways (RCP) can be used to provide further useful insights to reduce FS in the future. Furthermore, employing the highest-ranked ML models at higher resolution can also support developing interactive dashboards and decision support systems using open-source platforms (i.e., GEE), which can increase community engagement and capacity building enabling collective efforts to enhance disaster resilience. Finally, the results can further assist in developing future policies and decisions by focusing on the provided most optimal FSM ML algorithm. This will ultimately result in informed risk planning and adaptation, particularly in high-susceptibility zones.

6 | CONCLUSIONS

This study compared all relevant (14) ML models in terms of their reliability through accuracy assessment, and their scalability through prediction time over three cities of Sindh Pakistan, namely Shikarpur, Jacobabad, and Larkana—regions most affected in the recent floods of 2022. The results of this study demonstrate that accuracy-wise, LGBM and XGBoost stood at the top with adjusted accuracy of 0.93, and F1-score of 0.84 and 0.83, respectively. The estimated ROC and AUC rank LGBM, XGBoost, Random Forest, and gradient boosting algorithm at the top, each with a mean AUC of 0.84. In terms of scalability, XGBoost performed better with a prediction time of ~ 18 s compared to LGBM (22 s) and RF (31 s). Moving forward, we incorporate the explainable artificial intelligence (eAI) technique to provide insights into the importance of several flood conditioning factors in FS. This aspect adds an additional layer of interpretability to our results, making them more accessible and actionable for policymakers and stakeholders involved in flood risk management. This study presents compelling evidence of the accuracy and performance of ML models, with LGBM and XGBoost standing out as top performers, which can be adopted by the provincial and national disaster management authorities. Additionally, we highlight the scalability aspect, where XGBoost demonstrates superior prediction time, making it suitable for areas with computational constraints—reflecting the broader applicability of our work. Importantly, the evaluation framework presented in this study can be employed to prioritise and select the most optimal ML algorithm according to any specific region given that the FS is an in situ challenge and should be treated accordingly. The resultant reliable and accurate FS information based on the most optimal model could better inform decisions and policy production to cope with increasing flood risks in the future.

ACKNOWLEDGEMENTS

Sajjad M. is funded by the HKBU Research Grants Committee (Start-up Grant-Tier 1, RC-STARTUP/21-22/12) of the Hong Kong Baptist University, Hong Kong SAR. Waleed M. is supported by a postgraduate studentship from the HKBU Research Grant Committee (PhD studentship, 2022–2026).

CONFLICT OF INTEREST STATEMENT

The research is conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

DATA AVAILABILITY STATEMENT

We are thankful to all the institutes (mentioned within the text) for the provisioning of relevant data to carry out this valuable study. All the data used for several analyses are freely available and the resources are mentioned within the paper.

ORCID

Mirza Waleed  <https://orcid.org/0000-0003-0006-2490>

Muhammad Sajjad  <https://orcid.org/0000-0002-1576-1342>

ENDNOTES

- ¹ <https://www.pbs.gov.pk/content/brief-census-2017>
- ² <https://www.usgs.gov/>
- ³ <https://data.humdata.org/group/pak>
- ⁴ <https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-analyst/majority-filter.htm>
- ⁵ <https://www.esri.com/en-us/arcgis/products/arcgis-pro/overview>
- ⁶ numpy.org
- ⁷ https://scikit-learn.org/stable/supervised_learning.html

REFERENCES

- Adhikari, P., Hong, Y., Douglas, K. R., Kirschbaum, D. B., Gourley, J., Adler, R., & Robert Brakenridge, G. (2010). A digitized global flood inventory (1998–2008): Compilation and preliminary results. *Natural Hazards*, 55, 405–422. <https://doi.org/10.1007/s11069-010-9537-2>
- Akhtar, Z., Qazi, U., Sadiq, R., El-Sakka, A., Sajjad, M., Ofli, F., & Imran, M. (2023). Mapping flood exposure, damage, and population needs using remote and social sensing: A case study of 2022 Pakistan floods. <https://doi.org/10.1145/3543507.3583881>
- Al Daoud, E. (2019). Comparison between XGBoost, LightGBM and CatBoost using a home credit dataset. *International Journal of Computer Engineering*, 13, 6–10.
- Ali, G., Sajjad, M., Kanwal, S., Xiao, T., Khalid, S., Shoaib, F., & Gul, H. N. (2021). Spatial-temporal characterization of rainfall in Pakistan during the past half-century (1961–2020). *Scientific Reports*, 11, 6935. <https://doi.org/10.1038/s41598-021-86412-x>
- Amiri, A., Soltani, K., Ebtehaj, I., & Bonakdari, H. (2024). A novel machine learning tool for current and future flood susceptibility mapping by integrating remote sensing and geographic information systems. *Journal of Hydrology*, 632, 130936.
- Atif, S., Umar, M., & Ullah, F. (2021). Investigating the flood damages in lower Indus Basin since 2000: Spatiotemporal analyses of the major flood events. *Natural Hazards*, 108, 2357–2383. <https://doi.org/10.1007/s11069-021-04783-w>
- Devitt, L., Neal, J., Coxon, G., Savage, J., & Wagener, T. (2023). Flood hazard potential reveals global floodplain settlement patterns. *Nature Communications*, 14, 2801. <https://doi.org/10.1038/s41467-023-38297-9>
- Dodangeh, E., Choubin, B., Eigdir, A. N., Nabipour, N., Panahi, M., Shamshirband, S., & Mosavi, A. (2020). Integrated machine learning methods with resampling algorithms for flood susceptibility prediction. *Science of the Total Environment*, 705, 135983. <https://doi.org/10.1016/j.scitotenv.2019.135983>
- Eckstein, D., Künzel, V., & Schäfer, L. (2021). *The global climate risk index 2021*. Germanwatch.
- Fang, Z., Wang, Y., Peng, L., & Hong, H. (2021). Predicting flood susceptibility using LSTM neural networks. *Journal of Hydrology*, 594, 125734. <https://doi.org/10.1016/j.jhydrol.2020.125734>
- Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., Husak, G., Rowland, J., Harrison, L., Hoell, A., & Michaelsen, J. (2015). The climate hazards infrared precipitation with stations—A new environmental record for monitoring extremes. *Scientific Data*, 2, 150066. <https://doi.org/10.1038/sdata.2015.66>
- Gao, Y., Hasegawa, H., Yamaguchi, Y., & Shimada, H. (2022). Malware detection using LightGBM with a custom logistic loss function. *IEEE Access*, 10, 47792–47804.
- Ha, T. N., Lubo-Robles, D., Marfurt, K. J., & Wallet, B. C. (2021). An in-depth analysis of logarithmic data transformation and per-class normalization in machine learning: Application to unsupervised classification of a turbidite system in the Canterbury Basin, New Zealand, and supervised classification of salt in the Eugene Island minibasin, Gulf of Mexico. *Interpretation*, 9, T685–T710.
- Hawker, L., Uhe, P., Paulo, L., Sosa, J., Savage, J., Sampson, C., & Neal, J. (2022). A 30 m global map of elevation with forests and buildings removed. *Environmental Research Letters*, 17, 024016. <https://doi.org/10.1088/1748-9326/ac4d4f>
- Karakas, G., Kocaman, S., Gokceoglu, C., 2023. A hybrid multi-hazard susceptibility assessment model for a basin in Elazig Province, Türkiye. *International Journal of Disaster Risk Science*. <https://doi.org/10.1007/s13753-023-00477-y>
- Karra, K., Kontgis, C., Statman-Weil, Z., Mazzariello, J. C., Mathis, M., & Brumby, S. P. (2021). Global land use/land cover with sentinel 2 and deep learning. In *2021 IEEE international geoscience and remote sensing symposium IGARSS* (pp. 4704–4707). IEEE.
- Kaya, K., & Gündüz Ögüdücü, Ş. (2018). A binary classification model for PM10 levels. In *2018 3rd international conference on computer science and engineering (UBMK)* (pp. 361–366). Institute of Electrical and Electronics Engineers (IEEE). <https://doi.org/10.1109/UBMK.2018.8566285>
- Kumar, D., & Singh, O. (2024). Flood susceptibility mapping of the Markanda River basin in north-West India. In P. C. Pandey, R. Kumar, M. Pandey, G. Giuliani, R. K. Sharma, & P. K. Srivastava (Eds.), *Geo-information for disaster monitoring and management* (pp. 297–319). Springer International Publishing. https://doi.org/10.1007/978-3-031-51053-3_13
- Lehner, B., & Grill, G. (2013). Global river hydrography and network routing: Baseline data and new approaches to study the world's large river systems. *Hydrological Processes*, 27, 2171–2186. <https://doi.org/10.1002/hyp.9740>
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765–4774.
- Luu, C., Pham, B. T., Phong, T. V., Costache, R., Nguyen, H. D., Amiri, M., Bui, Q. D., Nguyen, L. T., Le, H. V., Prakash, I., &

- Trinh, P. T. (2021). GIS-based ensemble computational models for flood susceptibility prediction in the Quang Binh Province, Vietnam. *Journal of Hydrology*, 599, 126500. <https://doi.org/10.1016/j.jhydrol.2021.126500>
- Maxwell, A. E., Warner, T. A., & Guillén, L. A. (2021). Accuracy assessment in convolutional neural network-based deep learning remote sensing studies—Part 1: Literature review. *Remote Sensing*, 13, 2450.
- Mudashiru, R. B., Sabtu, N., Abustan, I., & Balogun, W. (2021). Flood hazard mapping methods: A review. *Journal of Hydrology*, 603, 126846. <https://doi.org/10.1016/j.jhydrol.2021.126846>
- Nkwunonwo, U. C., Whitworth, M., & Baily, B. (2020). A review of the current status of flood modelling for urban flood risk management in the developing countries. *Scientific African*, 7, e00269. <https://doi.org/10.1016/j.sciaf.2020.e00269>
- Notti, D., Giordan, D., Caló, F., Pepe, A., Zucca, F., & Galve, J. (2018). Potential and limitations of open satellite data for flood mapping. *Remote Sensing*, 10, 1673. <https://doi.org/10.3390/rs10111673>
- Olofsson, P., Foody, G. M., Herold, M., Stehman, S. V., Woodcock, C. E., & Wulder, M. A. (2014). Good practices for estimating area and assessing accuracy of land change. *Remote Sensing of Environment*, 148, 42–57. <https://doi.org/10.1016/j.rse.2014.02.015>
- Pandey, M., Arora, A., Arabameri, A., Costache, R., Kumar, N., Mishra, V. N., Nguyen, H., Mishra, J., Siddiqui, M. A., Ray, Y., Soni, S., & Shukla, U. (2021). Flood susceptibility modeling in a subtropical humid low-relief alluvial plain environment: Application of novel ensemble machine learning approach. *Frontiers in Earth Science*, 9, 1–26.
- Parsons, V. L. (2017). Stratified Sampling, in: Wiley StatsRef: Statistics Reference Online. pp. 1–11 <https://doi.org/10.1002/9781118445112.stat05999.pub2>
- Pham, B. T., Luu, C., Phong, T. V., Trinh, P. T., Shirzadi, A., Renoud, S., Asadi, S., Le, H. V., von Meding, J., & Clague, J. J. (2021). Can deep learning algorithms outperform benchmark machine learning algorithms in flood susceptibility modeling? *Journal of Hydrology*, 592, 125615. <https://doi.org/10.1016/j.jhydrol.2020.125615>
- Pradhan, B., Lee, S., Dikshit, A., & Kim, H. (2023). Spatial flood susceptibility mapping using an explainable artificial intelligence (XAI) model. *Geoscience Frontiers*, 14, 101625.
- Rentschler, J., Salhab, M., & Jafino, B. A. (2022). Flood exposure and poverty in 188 countries. *Nature Communications*, 13, 3527. <https://doi.org/10.1038/s41467-022-30727-4>
- Sajjad, M., Ali, Z., & Waleed, M. (2023). Has Pakistan learned from disasters over the decades? Dynamic resilience insights based on catastrophe progression and geo-information models. *Natural Hazards*, 117, 3021–3042. <https://doi.org/10.1007/s11069-023-05976-1>
- Schumann, G., Bates, P. D., Horritt, M. S., Matgen, P., & Pappenberger, F. (2009). Progress in integration of remote sensing-derived flood extent and stage data and hydraulic models. *Reviews of Geophysics*, 47, RG4001. <https://doi.org/10.1029/2008RG000274>
- Seleem, O., Ayzel, G., de Souza, A. C. T., Bronstert, A., & Heistermann, M. (2022). Towards urban flood susceptibility mapping using data-driven models in Berlin, Germany. *Geomatics, Natural Hazards and Risk*, 13, 1640–1662. <https://doi.org/10.1080/19475705.2022.2097131>
- Serdar, M. Z., Ajjur, S. B., & Al-Ghamdi, S. G. (2022). Flood susceptibility assessment in arid areas: A case study of Qatar. *Sustainability*, 14, 9792. <https://doi.org/10.3390/su14159792>
- Seydi, S. T., Kanani-Sadat, Y., Hasanlou, M., Sahraei, R., Chanussot, J., & Amani, M. (2023). Comparison of machine learning algorithms for flood susceptibility mapping. *Remote Sensing*, 15, 192. <https://doi.org/10.3390/rs15010192>
- Shah, S. A., & Ai, S. (2024). Flood susceptibility mapping contributes to disaster risk reduction: A case study in Sindh, Pakistan. *International Journal of Disaster Risk Reduction*, 108, 104503.
- Teng, J., Jakeman, A. J., Vaze, J., Croke, B. F. W., Dutta, D., & Kim, S. (2017). Flood inundation modelling: A review of methods, recent advances and uncertainty analysis. *Environmental Modelling & Software*, 90, 201–216. <https://doi.org/10.1016/j.envsoft.2017.01.006>
- Towfiqul, I. A. R. M., Talukdar, S., Mahato, S., Kundu, S., Eibek, K. U., Pham, Q. B., Kuriqi, A., & Linh, N. T. T. (2021). Flood susceptibility modelling using advanced ensemble machine learning models. *Geoscience Frontiers*, 12, 101075. <https://doi.org/10.1016/j.gsf.2020.09.006>
- Tsakiris, G. (2014). Flood risk assessment: Concepts, modelling, applications. *Natural Hazards and Earth System Sciences*, 14, 1361–1369. <https://doi.org/10.5194/nhess-14-1361-2014>
- Ullah, K., Wang, Y., Fang, Z., Wang, L., & Rahman, M. (2022). Multi-hazard susceptibility mapping based on convolutional neural networks. *Geoscience Frontiers*, 13, 101425. <https://doi.org/10.1016/j.gsf.2022.101425>
- Wahyuni, S., Sisinggih, D., & Dewi, I. A. G. (2021). Validation of climate Hazard group InfraRed precipitation with station (CHIRPS) data in Wonorejo reservoir, Indonesia. *IOP Conference Series: Earth and Environmental Science*, 930, 012042. <https://doi.org/10.1088/1755-1315/930/1/012042>
- Waleed, M., & Sajjad, M. (2022). Leveraging cloud-based computing and spatial modeling approaches for land surface temperature disparities in response to land cover change: Evidence from Pakistan. *Remote Sensing Applications: Society and Environment*, 25, 100665. <https://doi.org/10.1016/j.rsase.2021.100665>
- Waleed, M., Sajjad, M., Acheampong, A. O., & Alam, M. T. (2023). Towards sustainable and livable cities: Leveraging remote sensing, machine learning, and geo-information modelling to explore and predict thermal field variance in response to urban growth. *Sustainability*, 15, 1416. <https://doi.org/10.3390/su15021416>
- Waleed, M., Sajjad, M., Shazil, M. S., Tariq, M., & Alam, M. T. (2023). Machine learning-based spatial-temporal assessment and change transition analysis of wetlands: An application of Google earth engine in Sylhet, Bangladesh (1985–2022). *Ecological Informatics*, 75, 102075. <https://doi.org/10.1016/j.ecoinf.2023.102075>
- World Bank. (2022). Pakistan: Flood Damages and Economic Losses Over USD 30 billion and Reconstruction Needs Over USD 16 billion—New Assessment.

- World Meteorological Organization. (2022). *Provisional state of the global climate in 2022*. World Meteorological Organization.
- Zhang, Z., Tian, W., & Liao, Z. (2023). Towards coordinated and robust real-time control: A decentralized approach for combined sewer overflow and urban flooding reduction based on multi-agent reinforcement learning. *Water Research*, 229, 119498. <https://doi.org/10.1016/j.watres.2022.119498>
- Zhao, G., Pang, B., Xu, Z., Peng, D., & Xu, L. (2019). Assessment of urban flood susceptibility using semi-supervised machine learning model. *Science of the Total Environment*, 659, 940–949. <https://doi.org/10.1016/j.scitotenv.2018.12.217>

How to cite this article: Waleed, M., & Sajjad, M. (2025). 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*, 18(1), e13047. <https://doi.org/10.1111/jfr3.13047>