Application of machine learning algorithms for flood susceptibility assessment and risk management

R. Madhuri, S. Sistla and K. Srinivasa Raju

ABSTRACT

Assessing floods and their likely impact in climate change scenarios will enable the facilitation of sustainable management strategies. In this study, five machine learning (ML) algorithms, namely (i) Logistic Regression, (ii) Support Vector Machine, (iii) K-nearest neighbor, (iv) Adaptive Boosting (AdaBoost) and (v) Extreme Gradient Boosting (XGBoost), were tested for Greater Hyderabad Municipal Corporation (GHMC), India, to evaluate their clustering abilities to classify locations (flooded or non-flooded) for climate change scenarios. A geo-spatial database, with eight flood influencing factors, namely, rainfall, elevation, slope, distance from nearest stream, evapotranspiration, land surface temperature, normalised difference vegetation index and curve number, was developed for 2000, 2006 and 2016. XGBoost performed the best, with the highest mean area under curve score of 0.83. Hence, XGBoost was adopted to simulate the future flood locations corresponding to probable highest rainfall events under four Representative Concentration Pathways (RCPs), namely, 2.6, 4.5, 6.0 and 8.5 along with other flood influencing factors for 2040, 2056, 2050 and 2064, respectively. The resulting ranges of flood risk probabilities are predicted as 39–77%, 16–39%, 42–63% and 39–77% for the respective years.

Key words | flood risk, Hyderabad, hyperparameters, machine learning, RCPs

HIGHLIGHTS

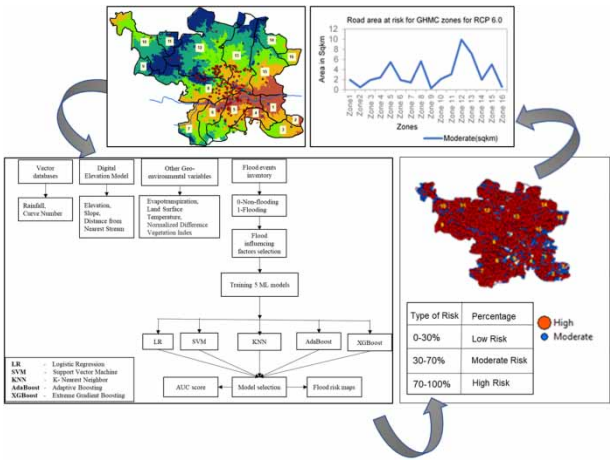
- Comparative assessment of ML algorithms to identify the most suitable algorithm for Greater Hyderabad Municipal Corporation (GHMC), India, to classify locations as either flooded or non-flooded.
- The most reliable ML algorithm (in this case XGBoost) is employed to predict flood risk probabilities for extreme rainfall situations in four different RCPs in association with other flood influencing factors.

R. Madhuri
K. Srinivasa Raju (corresponding author)
Department of Civil Engineering,
BITS Pilani Hyderabad Campus,
Hyderabad 500 078,
India
E-mail: *ksraju@hyderabad.bits-pilani.ac.in*

S. Sistla
Department of Chemical Engineering,
BITS Pilani Hyderabad Campus,
Hyderabad 500 078,
India

doi: 10.2166/wcc.2021.051

GRAPHICAL ABSTRACT



INTRODUCTION

Urbanisation, associated with intensified hydrological, ecological, environmental and climate changes, gives rise to reduction in groundwater recharge, evapotranspiration (ET) and infiltration, consequently resulting in an increase in quantity of runoff volume during most of the floods (Pirnia *et al.* 2019). Spatial and temporal characteristics of these parameters also impact the flooding area and flood risk is likely to escalate with the impact of climate change (Yin *et al.* 2015). With this, effective decision-making is possible with analytical and mathematical approaches by addressing climate change and its probable risks.

Complimentarily, the use of machine learning (ML) techniques in flood risk assessment is growing due to their ability to capture relationships efficiently (Wang *et al.* 2015) as the assessment depends primarily on detecting the flood-prone areas using historical events and topography (Costache & Zaharia 2017). Knowledge of flood-prone areas helps identify the locations that are susceptible to floods. ML algorithms were also applied in water resources and allied fields. Brief but relevant details are as follows.

Sarzaeim *et al.* (2017) used support vector machine (SVM), genetic programming (GP) and artificial neural networks (ANNs) for the prediction of runoff in climate change conditions for Aidoghmoush Basin, Iran. SVM outperformed GP and ANN by 7 and 5%, respectively, in terms of runoff. Modaresi *et al.* (2018) compared K-nearest neighbor (KNN) regression, ANN, least-square support vector regression (LS-SVR) and generalised regression neural network (GRNN) for their predictive ability of inflow for Karkheh dam, Iran. They concluded that ANN performs best. Kim *et al.* (2019) evaluated multilinear progression (MLP), multivariate adaptive regression splines (MARS), SVM and adaptive network-based fuzzy inference system (ANFIS) for Big Bend station, USA and found MLP to be superior. Wu *et al.* (2019) used extreme gradient boosting (XGBoost), Kernel-based nonlinear extension of Arps decline (KNEA), ANN, random forest (RF), MARS, gradient boosting decision tree (GBDT), SVM and extreme learning machine (ELM) for Jiangxi Province, China. The tree-based models were found to provide higher accuracy than others.

Ni *et al.* (2020) used XGBoost for streamflow forecasting in Cuntan and Hankou stations in China. XGBoost outperformed SVM in the case of Hankou station, while XGBoost was outperformed by SVM in the Cuntan station. Sankaranarayanan *et al.* (2020) employed SVM, KNN and Naïve Bayes for Bihar and Orissa for flood forecasting and compared them with deep neural networks (DNNs). Results showed that DNN provided better accuracy. Kaur *et al.* (2021) designed scheduling policies for power load using

AdaBoost, GBM (gradient boosting machine), XGBoost, RF and support vector classifier among which AdaBoost performed the best, slightly outperforming XGBoost. Oliveira and Carneiro (2021) created mineralogical models in the Santos Basin, namely, MLP, SVM, RF, AdaBoost and XGBoost. AdaBoost showed R^2 values above 0.9, outperforming XGBoost.

Ren *et al.* (2019) applied multiple linear regression, ANN, SVM and RF for Yarlung Zangbo River Basin, China, among which the RF model yielded the highest efficiency for downscaling the temperature for Representative Concentration Pathways (RCPs) of 2.6 and 8.5 for the period 2016–2050. Ahmed *et al.* (2020) employed ANN, KNN, relevance vector machine (RVM) and SVM to develop multi-model ensembles for precipitation, maximum and minimum temperatures for a case study in Pakistan using 36 global climate models (GCMs). It was observed that RVM and KNN demonstrated better skills. Hosseini *et al.* (2020) employed ANN, ANFIS and KNN for downscaling precipitation for Amameh station, Latyan dam basin, Iran. All methods showed satisfying results.

Tehrany *et al.* (2015) used various kernels in SVM to achieve prediction rates as high as 84.9% while assessing flood susceptibility in the Kuala Terengganu basin, Malaysia. Al-Abadi (2018) assessed flood susceptibility of a region in south Iraq using AdaBoost, RF and Rotation forest, among which AdaBoost gave the highest accuracy of 94.5%. Shahabi *et al.* (2020) performed flood susceptibility studies for the Haraz watershed in Iran, using various hybrid KNN models, of which a bagging model showed the highest area under curve (AUC) score of 0.811. López & Rodriguez (2020) used logistic regression (LR) to forecast flash floods in the city of São Paulo, Brazil and achieved an accuracy of 0.86. El-Haddad *et al.* (2021) employed four ML models to develop flood susceptibility maps for the Wadi Qena Basin, Egypt. The accuracy of ML models was evaluated using receiver operating characteristics (ROC) and the AUC. XGBoost proved superior to KNN with an AUC score of 0.902, for the flash flood prediction in Wadi El-Laqeita, Egypt (El-Magd *et al.* 2021). Gad & Hosahalli (2020) compared classification and predictive models using daily weather variables of stations in India and gathered from the National Climatic Data Center. They carried out 10-fold cross-validation to describe the AUC scores among other metrics. Islam *et al.* (2021) developed ensembling of ML models in the context of flood susceptibility in the Teesta sub-catchment, Bangladesh. AUC scores are based on a 10-fold cross-validation.

Julien *et al.* (2006) used harmonic analysis for the computation of normalised difference vegetation index (NDVI) and land surface ttemperature (LST) values to find out changes in vegetation in the European Continent between 1982 and 1999. Similar research work was initiated by Zhou *et al.* (2015) to calculate ET at 12 meteorological stations across Ontario, Canada. Wu *et al.* (2019) used RF, GBDT and XGBoost for the estimation of reference ET for the case study of China and those exhibited good estimation accuracy. Mosavi *et al.* (2021) used ensembled ML methods, namely, AdaBoost, Bagged classification and regression trees (Bagged CART), Boosted generalised additive model (GAMBoost) and RF for the Dezekord-Kamfiruz watershed, Iran to assess the groundwater potential. RF and Bagged CART performed better.

Research gaps and objectives

Very few studies were reported on AdaBoost and XGBoost for flood risk mapping in the international arena. However, no specific studies are reported in Indian conditions. Accordingly, this study is focused on Greater Hyderabad Municipal Corporation (GHMC), Telangana state, India. Thus, based on limited but effective research gaps, the objectives are framed as follows:

(i) Comparative assessment of ML algorithms, namely, (i) LR, (ii) SVM, (iii) KNN, (iv) Adaptive Boosting (AdaBoost), and (v) XGBoost for predicting flood susceptibility/identifying flooding locations for historic flood events and identifying the most suitable algorithm.
(ii) To compute future ET, LST and NDVI using harmonic regression.
(iii) To prepare maps of flood susceptible areas and predict flood risk probabilities for RCPs 2.6, 4.5, 6.0 and 8.5 in association with other flood influencing factors based on the suitable ML model (obtained from objective (i)).

Figure 1 presents a methodological workflow of the study. This paper describes the study area and data

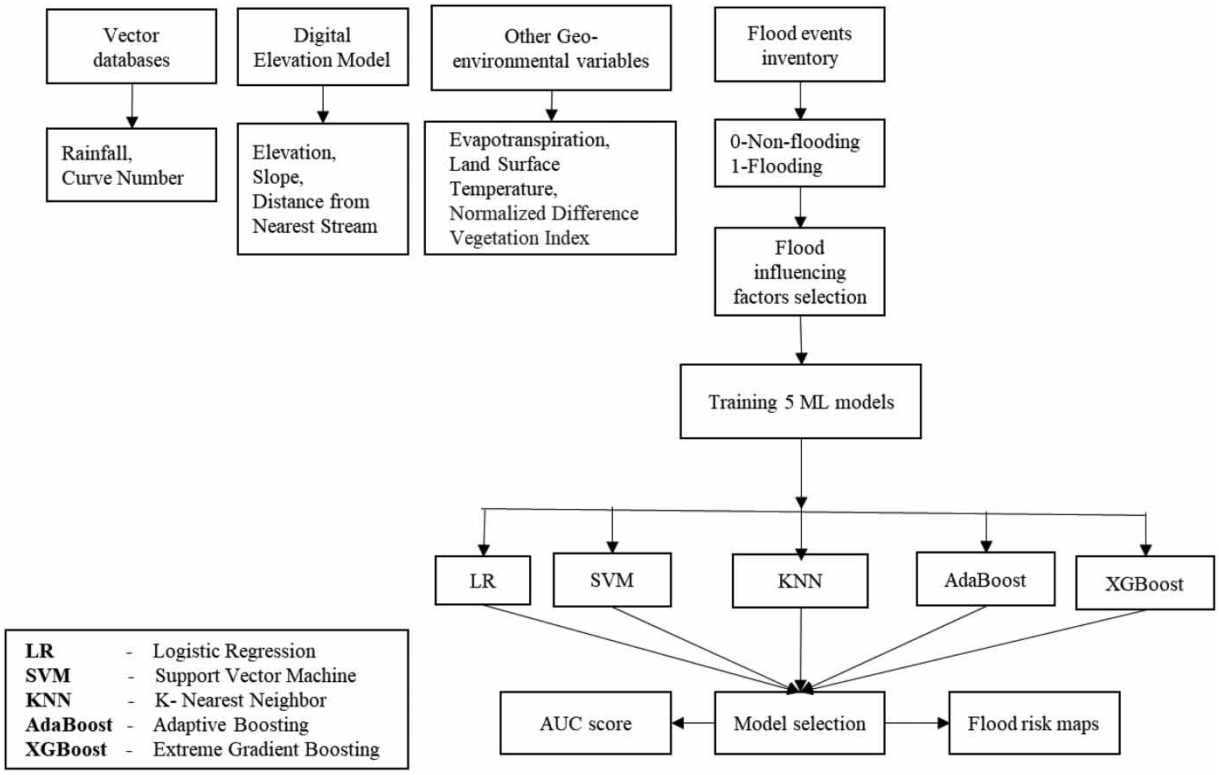


Figure 1 | Scheme of methodological workflow.

collection/processing, description of ML algorithms, discussion on presented results, summary and conclusions in the following sections.

STUDY AREA, DATA COLLECTION AND PROCESSING

The present study focuses on GHMC which covers an area of 625 km^2 and is divided into 16 zones as shown in Figure 2 based on the storm water drainage network. The Musi river runs in the middle of Hyderabad. Rainfall is the maximum in the monsoon season (June to October) which leads to heavy flash floods and submergence in the flood plains. Undulated terrain of GHMC, the low-lying regions especially near the Musi river have been the causes of flooding. The elevation in GHMC increases as we move away from the centre. Zones 1, 4, 5, 6 and 15 contain areas with lowest elevation (i.e. 462–506 m) and zones 2, 8, 12 and 13 contain areas with slightly higher elevation (506–531 m). With an increase in imperviousness, the city has suffered heavy losses due to high intensity urban flooding (Vemula *et al.* 2019). The flood locations for the years of extreme rainfall, i.e. 2000, 2006 and 2016, are also represented in Figure 2.

Factors influencing flooding

Finding the salient influencing variables which affect flooding is an important step to acquire flood susceptibility values (Rahmati *et al.* 2016). Flooded locations (dependent variables) were identified based on the available historical flood data of 2000, 2006 and 2016. Flooded and non-flooded locations are tagged values of '1' and '0', respectively, for the modelling purpose. Based on the literature, eight flood influencing factors (independent variables/inputs) were selected (Tehrany *et al.* 2015). These variables are rainfall, elevation, slope, distance from nearest stream (DNS), ET, LST, NDVI and curve number (CN).

Data collection relevant to historic and future studies is presented in the following sections.

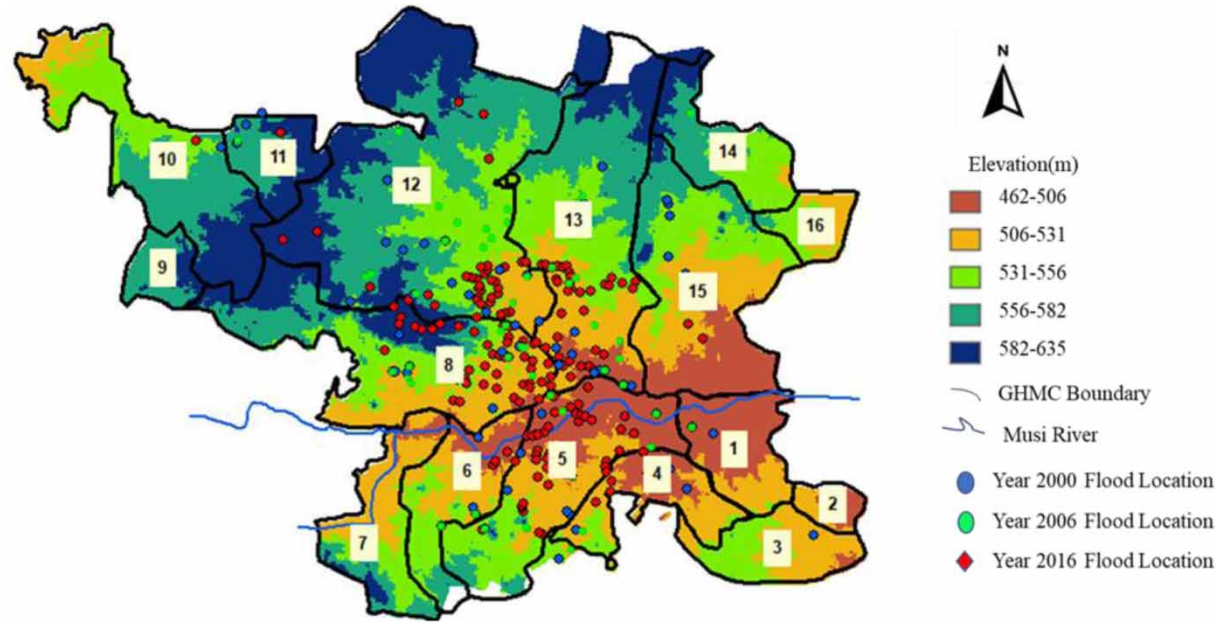


Figure 2 | Study area of GHMC showing flood locations.

Historic data

All the flood influencing factors in this study are processed pixel-wise. A pixel is the smallest unit of information on a raster map. Each pixel is a square cell containing the magnitude of the feature it refers to and of dimensions 30 m × 30 m for consistent resolution among the various features. The entire GHMC is contained in a box consisting of 1026 × 1421 such pixels. Pixel-wise data are collected for the years 2000, 2006 and 2016. The process of obtaining data for the year 2016 is described below. Similar processes were carried out for the years 2000 and 2006 to obtain the corresponding data.

- Rainfall data from 27 meteorological stations were procured from the Directorate of Economics and Statistics (DES) and GHMC. Each zone was assigned a rain gauge based on the closest meteorological station. The rainfall values were divided into five classes, i.e. from 1–35, 35–70, 70–105, 105–140 and 140–165 mm, respectively. Begumpet and Ameerpet (zone 12) meteorological stations have recorded the highest rainfall, and Serilingampally (zone 10) showed the least rainfall.
- Advanced Space-borne Thermal Emission and Reflection Radiometer (ASTER)-based digital elevation model (DEM) of 30 m × 30 m spatial resolution was utilised to create, extract and analyse the slope, elevation and streams of drainage basin (USGS 2016). The elevation data are derived from the DEM. Elevation ranges from 462 to 635 m and is divided into five classes as follows: from 462–506, 506–531, 531–556, 556–582 and 582–635 m (Figure 2). The areas near the Musi river, i.e. the central part of GHMC, have lower elevation as compared to the northern and southern parts of GHMC. The elevation data of GHMC are assumed to be constant over time.
- DEM analysis is used to calculate slope using the slope tool of spatial analyst in ArcGIS 10.1. The slope was classified into five groups between 0.3–1.22°, 1.22–2.29°, 2.29–4.28°, 4.28–7.95°, and 7.95°–19.82°. Only a small area in the middle of GHMC is completely flat, i.e. zero slope. This is because of the presence of a large water body: Hussain Sagar. High values of slope are seen in zones 5 and 1. As elevation of GHMC is assumed constant for all years, the slope data are naturally constant as well.
- LANDSAT USGS is Landsat data in the form of products (satellite imageries), which are held in the archives of USGS. They can be downloaded free from a variety of data portals. Land-use data are obtained from satellite imageries of LANDSAT, open street maps and GHMC.
- Stream data are obtained from the DEM using stream delineation of spatial analyst tool of ArcGIS 10.1. It is

possible to find the distance between a pixel and the closest stream to it. This process was iterated for all pixels in GHMC, and the DNS values were recorded. These were in the range of 0–2,130.5 m. Due to a dendritic structure of the streams spread across GHMC, there are very few pixels where extreme DNS values are present. The DNS values are assumed to be constant over time.

- ET values were obtained from the Moderate Resolution Imaging Spectroradiometer (MODIS) data from Google Earth Engine (GEE 2020). The data were converted from a 500 m × 500 m resolution to a 30 m × 30 m resolution using the reprojection Snap Raster tool of ArcGIS 10.1. This tool helps in snapping the raster cell size by matching it with their alignment (Feng *et al.* 2012). ET was split into classes of ranges 4–12.98 mm/day (very low), 12.98–21.35 mm/day (low), 21.35–31.88 mm/day (moderate), 31.88–50.78 mm/day (high) and 50.78–83 mm/day (very high) using ArcGIS 10.1. This classification is based on Jenks classification (Jenks 1967; Andualem & Demeke 2019).
- LST values were obtained from the MODIS data via the GEE. LST was split into classes of 30.27–32.97 °C (very low), 32.97–35.67 °C (low), 35.67–36.47 °C (moderate), 36.47–37.32 °C (high) and 37.32–39.21 °C (very high) using ArcGIS 10.1 using the Jenks classification scheme. The pixels obtained were of dimensions 1,000 m × 1,000 m and had to be reduced to 30 m × 30 m to make them uniform with the rest of the data obtained. This could be due to the urban sprawl and the urban heat island developed because of rapid urbanisation of the city.
- NDVI values were obtained from MODIS data via the GEE. The NDVI was classified in the ranges of 0.03–0.23 (very low), 0.23–0.29 (low), 0.29–0.35 (moderate), 0.35–0.42 (high) and 0.42–0.66 (very high). The pixels were reduced to 30 m × 30 m from 500 m × 500 m using the geographical information system (GIS). Zones 7 and 9 showed higher NDVI values than the others. This is because zone 9 is covered with grasslands, trees and plantation. Lower values are seen in the city due to urban agglomeration.

CNs are obtained using percentage imperviousness from land-use and soil data. Land-use data are obtained from satellite imageries of LANDSAT USGS and open street maps and GHMC. Soil data are obtained from the Directorate of Agricultural Commissionerate, Government of Telangana. Using this information, supervised classification is done for the years 1995 (as year 2000 data were not available), 2006 and 2016 to calculate percentage impervious land use from the obtained land use. This is used in association with the soil data to compute the CN for all pixels of GHMC using the CN grid in ArcGIS 10.1. The percentage imperviousness values of the years 1991, 2001 and 2015 were obtained from Sannigrahi *et al.* (2018) and Nayan *et al.* (2020). Imperviousness percentage for the year 2002 was interpolated between years 2001 and 2005. The Hyderabad Metropolitan Development Authority (HMDA) has proposed a master plan for the year 2031 (HMDA 2019) by the use of which imperviousness percentage was calculated. With these eight data points, the imperviousness percentage was fitted using a sigmoidal curve as shown in Equation (1). Sigmoidal fit had the best R^2 value of 0.96 among the various functions used and was thus selected for extrapolation. This also facilitates the interpolation of the value of the impervious percentage in the year 2000. The entire GHMC came under soil group C. Resulting CNs for impervious areas in 2000, 2006 and 2016 are 78, 83 and 85, respectively.

$$\text{Imperviousness percentage} = 90.48 - \frac{(65204330)}{\left(1 + \left(\frac{t}{1637.723}\right)^{72.44}\right)} \quad (1)$$

Here 't' refers to the year for which imperviousness percentage needs to be computed.

For example, the value of $t = 2{,}040$ gives an imperviousness percentage of 82.46.

The major flood events of the years 2000, 2006 and 2016 (numbering 68, 171 and 56 locations), respectively, were obtained from the GHMC Disaster Management Cell (GHMC 2018) and cross-checked using flood vulnerability maps obtained from the National Remote Sensing Centre (NRSC). Data augmentation is necessary before proceeding with ML algorithms since the number of data locations are 295 and are not sufficient. In order to solve this problem and to tackle the lack of non-flooded locations in the dataset, the

state (flooded or non-flooded) of an area flood-stricken in any of the three given years is considered. This is equivalent to taking a union of all the locations, and then each location's state along with all flood influencing factors corresponding to all three years is appended to the dataset. After performing this union, 71 locations where flooding is virtually impossible were also added to improve the separability of the data. This brings the dataset to a total of 1,086 locations × 8 features. Here, the 1,086 locations are obtained as follows: $1{,}086 = [(68 + 171 + 56) \times 3] - 12 + (71 \times 3)$, since 12 locations were the same in all the years considered. 28.6% of the 1,086 locations, i.e. 310 of them, are non-flooded, giving rise to a class ratio (ratio of the number of flooded and non-flooded locations) of 0.4.

FUTURE DATA GENERATION

To calculate the flood risks in future, it is essential to obtain values of the flood influencing factors for the future years. However, the slope, elevation and DNS of each pixel are assumed to remain constant in the future (the same value as considered for the year 2016).

To obtain the values of rainfall during extreme events in the years 2040, 2056, 2050 and 2064, GCM, GFDL-CM3 data by the nonlinear regression-based downscaling approach (Swathi 2020) were used. These years correspond to the highest rainfall events in RCPs 2.6, 4.5, 6.0 and 8.5. The values of rainfall obtained are as follows:

- 580.52, 579.67 and 580.44 mm in 2040 for RCP2.6 (23rd July–25th July; event is of 3 days)
- 415 mm in 2056 for RCP4.5 (2nd August–21st October; event is of 80 days; the rainfall values are not presented here due to lengthy data)
- 440.35 mm in 2050 for RCP6.0 (26th July–11th August; event is of 17 days; the rainfall values are 0.30, 7.21, 79.42, 117.61, 181.69, 35.39, 3.40, 2.33, 3.30, 2.71, 1.26, 0.88, 1.33, 1.99, 0.78, 0.57 and 0.18 mm)
- 624.2 mm in 2064 for RCP8.5 (24th July–11th August; event is of 19 days; the rainfall values are 0.21, 0.10, 1.29, 1.39, 0.83, 0.89, 2.82, 3.93, 5.27, 10.55, 17.78, 17.42, 15.70, 33.04, 48.59, 324.14, 135.86, 3.45 and 0.94 mm).

For forecasting the land imperviousness of the years 2040, 2050, 2056 and 2064, Equation (1) is used. Upon forecasting, it was found that the values of land imperviousness were expected to increase up to 79.71% in 2031, 82.47% in 2040, 84.86% in 2050, 85.93% in 2056 and 87.05% in 2064, respectively. Accordingly, estimated CNs for the years 2031, 2040, 2050, 2056 and 2064 were found to be 86, 87, 88, 88 and 89, respectively.

ET, LST and NDVI could also be calculated for the future using GCMs. However, resolution of the data is very low and not enough variation is obtained spatially in order to employ an ML algorithm successfully. Accordingly, the harmonic regression model is created for each pixel for ET, LST and NDVI, and these values could be predicted. Harmonic regression is a type of multiple regression model where trigonometric functions of a given variable, usually time, are taken for the predictor variables. It is used in the modelling of time-series datasets which are periodic in nature. The parameters such as ET, LST and NDVI can be predicted using harmonic analysis (Julien *et al.* 2006; Zhou *et al.* 2015; Wu *et al.* 2019) as follows (Equation (2)):

$$x_t = \mu + At + B\cos(2\pi t) + C\sin(2\pi t) \tag{2}$$

where μ is the mean of the time-series data, A is the co-efficient of the trend component, B and C are the amplitudes of variation of the cosine and sine components, respectively. This problem can be treated as multiple regression. The values of μ, A, B and C can be found/calibrated by multiple regression, and thus, values of the required features could be predicted.

For demonstrative purposes, a pixel in Begumpet, part of GHMC, is chosen, for which a harmonic model is constructed for ET, LST and NDVI (respectively, Equations (3)–(5)) from which values can be extrapolated. The ranges of R^2 values of the three regression models for all the pixels considered are 0.35–0.46, 0.42–0.57 and 0.36–0.46, respectively. Even though the harmonic regression models do not fit the data very precisely, the peaks and falls of the values of ET, LST and NDVI are captured by the model. Using this method, it is possible to roughly estimate what the values of above features will be in future decades. In Equations (3)–(5), t represents the year for which the corresponding feature is

being calculated.

$$\mathrm{ET} = 0.1 \times [115.63 + 5.80\,(t-2001) - 134.49\cos(2\pi(t-2001)) + 35.24^{*}\sin(2\pi(t-2001))] \quad (3)$$

$$\mathrm{LST} = 0.02 \times [14994.01 - 1.72(t-2000.26) - 334.08\cos(2\pi(t-2000.26)) + 497.79^{*}\sin(2\pi(t-2000.26))] \quad (4)$$

$$\mathrm{NDVI} = 0.0001 \times [158.467 + 19.71(t-2000.21) - 2055.54\cos(2\pi(t-2000.21)) + 612.28^{*}\sin(2\pi(t-2000.21))] \quad (5)$$

Equations (3)–(5) obtained through harmonic regression were validated for comprehensive understanding. The mean of NRMSE values (taken across all pixels), respectively, for ET, LST and NDVI between observed and simulated are 0.307, 0.230 and 0.305. The mediocre values of NRMSE may be attributed to the model's inability to predict the extreme peaks of the features. The points which are outliers in the harmonic regression can be considered inconsequential due to their sparse nature.

Having developed harmonic regression models for three features, it is possible to estimate future values using the models created, similar to the examples given above. Future values of ET, LST, and NDVI are extrapolated for 2040, 2056, 2050 and 2064 for Begumpet (for demonstration) using Equations (3)–(5) as follows:

- ET: 404.33, 402.47, 520.96 and 433.73 kg/m^2/8 days;
- LST: 310.18, 305.149, 308.72 and 308.98 K;
- NDVI: 0.30, 0.30, 0.35 and 0.32.

In a similar process, the data were extrapolated for each location in GHMC. Extrapolated results were used to populate the databases of the future years. Then, the selected ML algorithm was used to generate flood risk maps for 2040, 2056, 2050 and 2064 under the most extreme rainfall cases. The next section presents ML algorithms.

Machine learning algorithms

A number of ML algorithms can be used, and each algorithm works differently for the dataset it is being fitted to. The problem tackled in this study is referred to as supervised binary classification. For optimal fitting, the hyperparameters should be chosen carefully. The hyperparameters chosen by tuning are reported in Supplementary Table S1 (presented in appendix).

Logistic regression

Binary LR is a supervised classification approach used for predicting classes of dichotomous nature ('0' or '1'). It uses the logistic function (range [0,1]) to make predictions (Pradhan 2010). The output of LR is the probability that a particular data point belongs to the '1' or positive class (Equation (6)):

$$P = \frac{1}{1+e^{-z}} \quad (6)$$

If the probability output of LR is greater than a certain threshold, the point is classified as '1'; if lesser, it is classified as '0'. The inputs are $x_1, x_2, x_3 \ldots x_n$, and presented as follows in the following equation:

$$z = w_0 + w_1x_1 + w_2x_2 \ldots + w_nx_n \quad (7)$$

The weights, $w_0, w_1\, w_2\, w_3 \ldots w_n$, are obtained to give the best separation between the classes in the training data examples. A threshold is chosen to give a true positive rate (TPR) or false positive rate (FPR) acceptable for the situation.

Support vector machine

SVM is a supervised classification algorithm, which classifies datasets using hyperplanes in a high- or infinite-dimensional space. Each data point is classified as either '1' or '-1' depending on which side of the hyperplane it lies. A hyperplane can be defined as in the following equation:

$$w^Tx + b = 0 \quad (8)$$

where w is a vector of weights and b is the bias term. These parameters are trained allowing for some misclassification of the training data so as not to risk overfitting. A slack variable ξ is introduced for each data point, representing how far away the misclassified point is from the hyperplane ($\xi = 0$ for correctly classified points). The optimisation simplifies

finding out the minimum as in the following equation:

$$\min\left(\frac{1}{2}\ \|w\|_2 + C\ \sum \xi_n\right) \tag{9}$$

$$\text{s.t}\ y_n[w^T x + b]\ \geq 1 - \xi_n\ \forall\ n$$

$$\xi_n \geq 0\ \forall\ n$$

where n ranges over all the data points, C is the hyperparameter which decides how 'lenient' the classifier is, and y_n is the corresponding class of the data point. However, just the data point x, more often than not, is not linearly separable by a hyperplane. It becomes necessary to map x to a more complex, nonlinear space, say $\varnothing(x)$. The process of mapping data to higher dimensions can prove computationally expensive, and thus, the kernel trick is used.

The kernel trick reduces the computational time required for SVMs (Raghavendra & Deka 2014) by avoiding computation that transforms the data from low to high dimensions, and calculates relationships in the infinite dimensions used by the radial kernel directly (Equation (10).

$$K(x, x') = \exp\left(\frac{-||x - x'||^2}{2\sigma^2}\right) = \Phi^T(x)\ \Phi(x') \tag{10}$$

where x and x' are any two data points. All the inner product terms $\Phi^T(x)\Phi(x')$ that appear in the optimisation, which are computationally intensive, are substituted by the kernel function K. In the radial basis function (RBF) kernel (selected via hyperparameter tuning), the amount of influence one observation has on another is a function of the squared distance. The parameter σ defines how much influence one point has on another. The RBF kernel helps the conversion of data to infinite dimensions so as to make the linearly unclassifiable data classifiable. Hyperparameters related to SVM are presented in Supplementary Table S1.

K-nearest neighbors

KNN is a supervised, classification algorithm that works by first storing all the data and classifying new data points based on distance functions with respect to the stored data. Distance from the testing point to each training point is calculated as the following equation:

$$\text{Distance} = (x_{\text{train}} - x_{\text{test}})^p \tag{11}$$

where p is the Minkowski metric ($p \in [1, 2, 3, 4]$), x_{train} is a training data point and x_{test} is the data point whose class we wish to predict.

Then, among the 'K' nearest neighbors to the testing point, the class which is most common is assigned to the training point. Hyperparameters related to KNN are presented in Supplementary Table S1.

Adaptive boosting

Boosting refers to ensemble methods which use several weak learners to make a strong learner. The class of the ith training example is denoted as $y_i(-1$ or $+1)$. AdaBoost is an ML meta-algorithm, which selects features that improve the prediction of a model. It makes an ensemble out of 'weak' learning algorithms (for example, decision trees) to increase performance (Freund & Schapire 1997). In this implementation, decision trees are chosen as the weak learners. At each step, the 'hardness' of each training data point is input to the algorithm, so that newer generated trees classify the 'tougher' ones.

Here, the weak learner is a decision tree with two leaves, called a 'Stump'. Some stumps have a higher weightage in predicting the data, and each subsequent stump rectifies the errors of the preceding stump. The final prediction is given as in the following equation:

$$H(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right) \tag{12}$$

where $h_t(x)$ is the prediction by the tth weak classifier, α_t is the weight of the tth classifier and α_t is computed as in the following equation:

$$\alpha_t = \frac{1}{2}\ln\left(\frac{1 - \epsilon_t}{\epsilon_t}\right) \tag{13}$$

where ϵ_t is the fraction of misclassification by the tth classifier.

Each sample in the dataset is assigned a weight $D_t(i)$, which is initially equal, whose sum is unity. The feature, which differentiates between the classes the most, is chosen as the first stump, and based on its accuracy, the weightage of the stump is decided. The wrongly classified sample weights are increased as in the following equation:

$$D_{t+1}(i) = \frac{D_t(i)\exp(-\alpha_t y_i h_t(x_i))}{\sum_i D_t(i)\exp(-\alpha_t y_i h_t(x_i))} \tag{14}$$

Then, another dataset of the same size of the original dataset is made, randomly picked with repetition of samples from the previous dataset. Samples whose weights are higher get picked more often. Then, on this bootstrapped dataset, another iteration is carried out, and so on. Hyperparameters related to AdaBoost are presented in Supplementary Table S1.

XGBoost

XGBoost uses tree pruning and handles missing values. It minimises an objective function which is regularised, combined with a penalty term for model complexity. XGBoost trees are fit to the residual probabilities (Wu *et al.* 2019). Examples are first grouped in such a way that similar residuals are in the same group, and then branched off. Equation (15) provides a similarity score of each threshold:

$$\text{Similarity score} = \frac{(\sum \text{Residual}_i)^2}{\sum [p'_i \times (1 - p'_i)\,] + \lambda} \tag{15}$$

where λ is the regularisation parameter and p'_i is the previous probability computed for the ith training example in that branch.

We must also keep in mind the cover (min_child_weight), the minimum number of residuals in each leaf. Usually, the minimum value for cover is set to 1. Cover is calculated as $\sum [p'_i \times (1 - p'_i)\,]$. If the cover is less than minimum, the leaf is removed. The tree pruning is facilitated with the Tree Complexity Parameter. Equation (16) provides output values (w) for all leaves:

$$w = \frac{\sum \text{Residual}_i}{\sum [p'_i \times (1 - p'_i)\,] + \lambda} \tag{16}$$

The probabilities are updated in terms of logs of odds. The output value is multiplied by the learning rate and subtracted from the previous probability, and also expressed in log of odds terms. The probability after one such iteration can be obtained by converting the obtained value back to probability. Ultimately, XGBoost tries to minimise the objective function in the following equation:

$$O(y_i, p_i, w) = \sum_{i=1}^{n} L(y_i, p_i) + \frac{1}{2}\lambda w^2 \tag{17}$$

All the above equations and expressions were derived taking the loss function, $L(y_i, p_i)$, as the negative log-likelihood function in the following equation:

$$L(y_i, p_i) = [-y_i \log(p_i) + (1 - y)\log(1 - p_i)] \tag{18}$$

More details about XGBoost and related approaches are available (Ni *et al.* 2020; Kaur *et al.* 2021; Oliveira & Carneiro 2021). Hyperparameters related to XGBoost are presented in Supplementary Table S1. The next section presents the Results and Discussion.

RESULTS AND DISCUSSION

Hyperparameter tuning based on historic data

Usually, the entire dataset (1,086 locations × 8 features) is split into three subsets, the 'train' dataset to train the ML models, the 'validation' dataset, used for hyperparameter tuning, and the 'test' dataset, used for evaluating the model's performance. However, it is difficult when data are limited. Thus, in order to tackle this problem, the dataset will be used to validate itself via K-fold cross-validation. The performance metrics obtained through K-fold cross-validation will be used to assess the ML models. The hyperparameter tuning itself is carried out by using Bayesian optimisation. Bayesian optimisation assists in picking a set

of optimum hyperparameter values, which result in the model providing the best classification, based on the area under receiver operating characteristic curve (AUROC/AUC).

To boost performances further, recursive feature elimination using cross-validation (RFECV) was also performed to get the subset of flood influencing factors (features for ML algorithms) which provided the best AUC score for each algorithm.

We adopted 10-fold cross-validation in our study based on satisfactory results in the respective research papers (presented in the literature review section). Before applying any of the ML algorithms, the data are standardised normally, as per: $x^* = ((x - \mu)/\sigma)$, where x is an input variable, and μ and σ are mean and standard deviations of the corresponding feature.

Figure 3(a) presents the mean ROC curves for all ML algorithms considered. The means were calculated by finding out the ROC curve of each fold in the 10-fold cross-validation performed, and their means were reported along with the standard deviations (as $\mu \pm \sigma$) of each fold for each algorithm. The hyperparameters of all models are tuned as shown in Supplementary Table S1, and the resulting AUC scores are presented in Figure 3(a). XGBoost is marginally better than AdaBoost on the training data (AUCs of 0.83 and 0.82, respectively), and both significantly outperformed KNN, SVM and LR (AUCs 0.77, 0.74 and 0.71). XGBoost and AdaBoost also showed lesser standard deviation compared to other algorithms.

Validation

An ML model ideally should not be evaluated on the same dataset it was trained on, since the model will over-fit the dataset, and the results will be slightly better than they would have been. Hence, in order to get an unbiased evaluation of the model, it is necessary to run it on a test dataset that the model has not previously used.

Therefore, a dataset constructed for the year 2020 was used, and 16 locations afflicted by floods were obtained. Due to lack of spatially varying rainfall data for the year 2020, the entirety of GHMC was assumed to have the same rainfall of 674.22 mm for the extreme flood event, which is obtained from the meteorological station in Begumpet. The elevation, slope and DNS for each location are assumed to be the same as they were in the past at each location. The impervious land-use data are 77% which is computed using Equation (1) and accordingly CN is 85. ET, LST and NDVI values are obtained from the MODIS data using the GEE.

In order to maintain an analogous class ratio between testing and training data, seven non-flooded locations were also added to the dataset, making a total of 25. This ensures that the class ratio of the testing set matches that of the training set which was 0.4. The AUC curves and the mean AUC scores are shown in Figure 3(b).

The results using the various algorithms on the testing data show that XGBoost has an AUC score of 0.83, followed by AdaBoost, KNN, SVM and LR with AUC scores of 0.74, 0.66, 0.65 and 0.54, respectively. Thus, XGBoost performed at almost identical levels in both the training and validation,

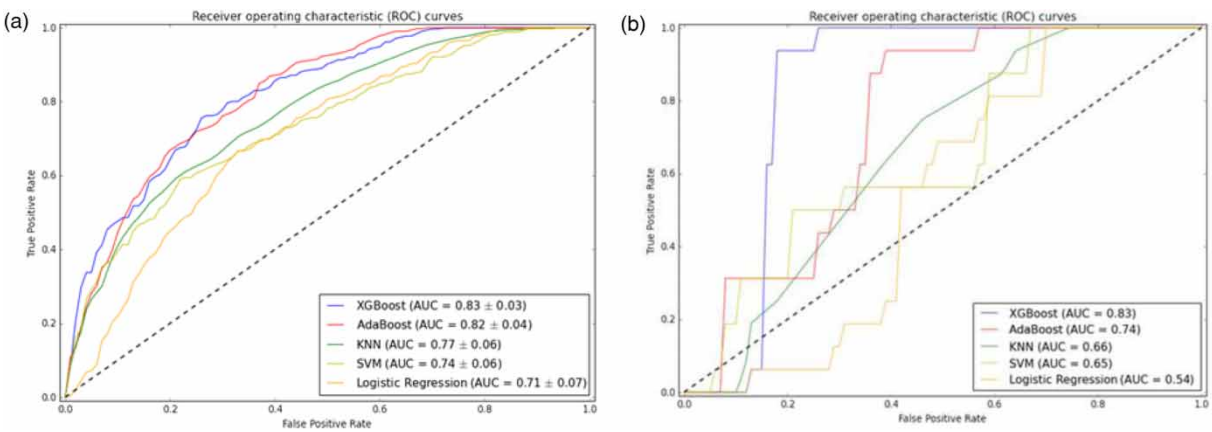


Figure 3 | (a) AUC training curves for chosen machine learning algorithms. (b) AUC curves for validation year 2020 for chosen machine learning algorithms.

unlike AdaBoost, which suggests that the model is neither overfitting nor under fitting, thus being a suitable model to choose. Hence, XGBoost was preferred to simulate the flood risk probabilities of future scenarios, and our training philosophy of selecting XGBoost was found to be valid. Since TPR is high compared to FPR, an appropriate classifier can be the one which appears at FPR of 0.14 and TPR of 0.94.

To illustrate the construction of AUC curves, let us consider the above point. At that probability threshold, 17 out of the 18 flood locations are predicted correctly, giving us a TPR of 0.94 (17/18), while six out of seven non-flooded locations are predicted correctly, giving us an FPR of 0.14 (1/7). Similar calculations are carried out for every classifier threshold level.

Analysis of flood maps from XGBoost

So far in the study, ML algorithms have been used to predict the probability of flooding at various discrete locations where floods have occurred in the past. After identifying a suitable ML algorithm (XGBoost), it can be used to map every location in GHMC to predict its corresponding flood probability. We used extreme rainfall based on RCPs. We have taken ET, LST, NDVI and CN (which are independent of RCPs) projected for the corresponding RCP years. Hence, we used terminology of years instead of RCPs for uniformity in the problem. After obtaining future datasets for 2040, 2056, 2050 and 2064, XGBoost is applied to them to obtain raw probability data for each point in GHMC. The results are depicted in Supplementary Figure S1.

The raw probability data obtained from XGBoost as seen in Supplementary Figure S1 (presented in appendix) are then assessed through the GIS using ArcGIS 10.1. This acts as the base layer/map where the features of roads, buildings and flood influencing factors are overlaid to correlate them with flood probability. The results in Supplementary Figure S1 show that the flood risk probabilities for the years 2040, 2056, 2050 and 2064 vary from 39–77% (moderate to high risk), 16–39% (low to moderate risk), 42–63% (moderate risk) and 39–77% (moderate to high risk), respectively. Here, flood risk area is based on the number of flood pixels.

In this assessment, roads and buildings are considered separately for susceptibility studies as they are more prone to damage and can cause devastation to the catchment during flooding. After analysing various areas of roads and building under risk, the elevation, slope, ET, LST, NDVI and DNS data (in pixel form) are overlapped with the probability maps obtained from XGBoost. The results from the above analysis are further used to study what effect the various ranges of flood influencing factors have on flood risk.

After superimposing the data of roads and buildings on the flood probability maps, the breakup of areas of each zone, depending on how much of their roads and buildings fall under high, moderate and low risks, respectively, are obtained. This breakup is reported for the years 2040, 2056, 2050 and 2064. This information shows a detailed analysis of the flood risk occurring each year. Supplementary Table S2 (presented in appendix) shows the percentage of buildings and roads (in the entire GHMC) under high, moderate and low risks. It may be noted that the area of roads in GHMC totals up to 50.64 km^2, and the total area of buildings in GHMC sums up to 84.91 km^2. Based on these, it is possible to calculate the percentage of roads and buildings under risk in the four future years considered, as given in the last two sections of Supplementary Table S2. These tables help visualise a broader picture, providing an overview of the possible flooding situations that can occur in the future.

A visualisation of the above values of probability risk is shown in Supplementary Figure S2 (presented in appendix). Each circle represents a building, and its colour represents the class of risk it falls into. It gives an overall view of the distribution of buildings that are under risk. Using Supplementary Table S2 and Figure S2 as references, several observations can be made for each year in the future summarised as follows:

- In the year 2040, zone numbers 2, 3, 5, 7, 8, 12, 14 and 16 have over 70% of their buildings and roads under high risk. Another alerting observation is that more than three quarters of the roads under zone 1 are under high risk. This may be attributed to the fact that areas in these zones lie at an elevation of less than 500 m. The maximum value of DNS for zone 2 is 991.45 m, followed by zone 5 with 1,202.31 m, which is comparatively lower than the rest of the zones. It is observed that the mean

slope of zone 12 is around 7.55°, which is comparatively higher than other zones that create a flooding situation in low-lying areas.

- In 2056, none of the zones are under high risk. Most of the zones have almost all of their area under low risk, with the exceptions of 10 and 11. In 2056, the roads of zone 12 are most affected, with over 34% of its roads under moderate risk. This can be attributed to the low intensity rainfall (0.7–20 mm daily), which occurs over a long duration of 80 days (2nd August–21st October 2056). This leads to fewer chances of water accumulation and stagnation. The areas with highest slopes (zones 10 and 11) observe maximum values of flood risk in this year.
- In the year 2050, all roads and buildings are under moderate risk. This may be due to the high intensity rainfalls on 29th July and 30th July 2050 (whose magnitudes are very high, i.e. 117 and 181 mm) with the otherwise continuous moderate rainfall from 26th July to 11th August 2050. This lack of continuous and significantly high rainfall puts virtually all locations under moderate risk, with an increase of risk in low-lying areas. No zones are under high risk in this year as well.
- In 2064, much like 2040, zone 2's buildings are mostly under high risk, while also having the highest percentage area of roads (0.44 out of 0.49 km^2, i.e. 90%) under high risk. In all zones, a large percentage of areas under high risk, with zones 1, 2, 3, 5, 7, 8, 12, 13, 14 and 16 have more than 70% of their buildings. Hence, due to the maximum percentage of buildings in zones 12 and 13, high risk is observed in these zones. Flooding may arise in these situations where the CN has a very high value of 89 or more, the elevation being the dominant factor which also causes inundation in the low-lying areas. Zones 2, 9 and 16 have no storm water network.

A similar analysis was done for the features of elevation, slope, ET, LST, NDVI and DNS. To better understand the direct dependence of factors on flooding, the ranges and means of the flood influencing factors corresponding to each range of flood risk probability were calculated, and inferences were drawn as follows:

- For the year 2040, the highest flood risk probability (70–80%) was the most common, occuring in 38.9% of the area. The lowest flood risk is found in the small range of 525–613 m, while the highest flood risk is found in the much wider range of 462–635 m. With respect to slope, areas with a slope of 15.79° are more vulnerable to flooding than areas with a higher slope of 19.82°. In 2040, GHMC expected to receive high intensity rainfall of 580 mm (daily) over a very short duration of 3 days. In such rainfall circumstances, having too high a slope will lead to high runoff, and water is not given a chance to accumulate. Similarly, in areas having low slope, water does not flow as much and there is not much contribution of water from nearby areas. A moderately high slope of 15.79° may lead to conditions where both accumulation and a slower rate of runoff occur, thus resulting in flooded areas. With respect to ET, areas under high risk (above 60% probability) lie in the first class of ET (4–12.98 mm/day). It was observed that the high-risk probabilities (above 70%) are observed in areas where the minimum value of NDVI is below 0.05. The number of pixels under high flood is 95.
- In 2056, most flood risk probabilites are concentrated between 20 and 30%. The least amount of flooding is in locations with less than 12° slope and mean elevation 567 m. In this year, a mean DNS of almost 470 m is observed in the areas with the highest risk, while a mean DNS of almost 400 m is observed in the areas with the lowest risk. This further reinforces the hypothesis that flooding increases as DNS increases. The rainfall in year 2056 is of low intensity (daily rainfall of range between 0.7 mm to 20 mm) and has a duration of 80 days. Given the low intensity, long duration of rainfall, and dendritic structure of streams, it is possible that floods will be well controlled. A steady increase of flood risk probability (10–20%, 20–30% and 30–40%) is observed with decline in the means of the ET (22.37, 19.41 and 18.54 mm/day) and decline in the means of NDVI (0.64, 0.60 and 0.54), respectively. The number of pixels under high flood is 32.
- For the year 2050, two-thirds of the total area comes under the flood risk probability range of 50–60%. Contrary to the year 2040, an increase in flood risk probability is observed with a decrease in slope, and the highest flood risk areas (60–70%) have elevations in the relatively narrow range of 518–616 m. 66% of the area under risk is at a mean DNS of 447 m (comparatively highest), the farthest

location being 1,938 m from the nearest stream. The rainfall magnitude is moderate (0.17–180 mm daily) and for a relatively longer duration (26th July to 11th August 2050). Hence, the moderate risk areas are more prominent as compared to high-risk areas in all the zones. When there is a prolonged consistent rainfall, the soil gets saturated, promoting flooding. Flood risk probabilities of over 50% are seen in areas with the NDVI greater than 0.3 on average. The number of pixels under high flood is 153.

- For the year 2064, one-third of the total area comes under the flood risk probability of 60–70%, while another third is under 70–80%. The highest flood risk probability range (70–80%) occurs at a very low elevation of 540 m. Similar to the year 2040, the maximum flood risk probability corresponds to a moderately high value of slope (12.4°) and a maximum DNS of 2 km. The rainfall in 2064 is continuous and of high intensity: from 24th July to 11th August 2064. The highest magnitude is 324 mm followed by 135 mm on 8th and 9th August 2064, respectively. The continuous rainfall provides less time for the catchment to capture all the rainwater in natural streams, causing floods in low-lying areas. The means of ET steadily increase with a rise in flood risk probability (40–80%). The means of NDVI corresponding to the flood risk areas of 30–80% first decrease and then stabilise with an increase in the flood risk probability. The number of pixels under high flood is 132.

SUMMARY AND CONCLUSIONS

In this study, five ML algorithms, namely, LR, SVM, KNN, AdaBoost and XGBoost, were applied and their performances were evaluated. They were used for the flood susceptibility mapping of GHMC, India, which has seen severe damage to property due to flooding. The training dataset was augmented by taking a union of the locations of the flooded regions in the years 2000, 2006 and 2016, and the features for the aforementioned locations in the corresponding year were collected accordingly. The models were both trained and hyperparameter tuned on this dataset using 10-fold cross-validation. Each model also underwent RFECV for the best classification. XGBoost had the best discriminatory behaviour with an AUC score of 0.83, followed by AdaBoost.

The trained and tuned models were tested on a dataset constructed on the flooded locations of the year 2020 of GHMC. This indicated that among the algorithms chosen, XGBoost is the most suitable algorithm to perform flood susceptibility mapping for GHMC. Then, XGBoost was used to generate flood risk maps under the most extreme rainfall cases for years 2040, 2056, 2050 and 2064. The prediction of flood locations for these years provides a clear view of how to take mitigation measures and also how to have an early preparedness for any calamitous situations.

Hence, it can be concluded that GHMC is mostly in the high-risk probabilities for year 2064. This is because the duration of flood is higher as compared to extreme rainfall in the remaining years. It can be inferred from analysis that year 2050 has moderate flood risk probability.

The novelty of this study lies in applying for the first time:

- Flood susceptibility mapping using ML methods in association with the GIS was generated more rigorously; and
- First application of XGBoost in modelling flood susceptibility maps.

This work is expected to set a benchmark for future scenarios to help authorities identify high-risk locations so as to undertake pre-emptive measures, duly accounting for dynamic changes and all uncertainties in the variables considered. However, our approach was limited by the lack of future rainfall data across all locations, currently using the projected rainfall values at only one location. The study can be improved by having future rainfall which varies spatially, including climate variables considered by predicting future ET and LST. The analysis carried out here is based on a mathematical approach, through which the occurrence of flood and non-flooding results can be obtained. However, by applying conceptual models such as HEC-RAS and HEC-HMS, the extent of flooding with the amount of flood depth can also be predicted with dynamic field data which is proposed as a future study.

ACKNOWLEDGEMENTS

This work is supported by Information Technology Research Academy (ITRA), Government of India under ITRA-water grant ITRA/15(68)/water/IUFM/01. The third

author would like to acknowledge the funding support provided by the Council of Scientific and Industrial Research (CSIR), New Delhi through a project with reference number 22(0782)/19/EMR-II dated 24 July 2019. The authors thank Officials of GHMC and other agencies for their help. Acknowledgements to Prof D. Nagesh Kumar, Department of Civil Engineering, Indian Institute of Science, Bangalore for providing valuable suggestions for improving the manuscript. The first author is thankful to Dr Swathi Vemula for providing future rainfall data and valuable discussion.

CONFLICT OF INTEREST

None.

DATA AVAILABILITY STATEMENT

Data cannot be made publicly available; readers should contact the corresponding author for details.

REFERENCES

Ahmed, K., Sachindra, D. A., Shahid, S., Iqbal, Z., Nawaz, N. & Khan, N. 2020 Multi-model ensemble predictions of precipitation and temperature using machine learning algorithms. *Atmospheric Research* **236**, 104806.

Al-Abadi, A. M. 2018 Mapping flood susceptibility in an arid region of southern Iraq using ensemble machine learning classifiers: a comparative study. *Arabian Journal of Geosciences* **11**, 218.

Andualem, T. G. & Demeke, G. G. 2019 Groundwater potential assessment using GIS and remote sensing: a case study of Guna Tana landscape, upper blue Nile Basin, Ethiopia. *Journal of Hydrology: Regional Studies* **24**, 1–13.

Costache, R. & Zaharia, L. 2017 Flash-flood potential assessment and mapping by integrating the weights-of evidence and frequency ratio statistical methods in GIS environment – case study: Bâsca Chiojdului River catchment (Romania). *Journal of Earth System Science* **126** (4), 59.

El-Haddad, B. A., Youssef, A. M., Pourghasemi, H. R., Pradhan, B., El-Shater, A. H. & El-Khashab, M. H. 2021 Flood susceptibility prediction using four machine learning techniques and comparison of their performance at Wadi Qena Basin, Egypt. *Natural Hazards* **105** (1), 83–114.

El-Magd, S. A., Pradhan, B. & Alamri, A. 2021 Machine learning algorithm for flash flood prediction mapping in Wadi El-Laqeita and surroundings, Central Eastern Desert, Egypt. *Arabian Journal of Geosciences* **14** (4), 1–14.

Feng, M., Huang, C., Channan, S., Vermote, E. F., Masek, J. G. & Townshend, J. R. 2012 Quality assessment of Landsat surface reflectance products using MODIS data. *Computers & Geosciences* **38** (1), 9–22.

Freund, Y. & Schapire, R. E. 1997 A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences* **55**, 119–139.

Gad, I. & Hosahalli, D. 2020 A comparative study of prediction and classification models on NCDC weather data. *International Journal of Computers and Applications*. https://doi.org/10.1080/1206212X.2020.1766769.

GEE 2020 Google Earth Engine. Available from: https://earthengine.google.com.

GHMC Disaster Management Cell 2018 Available from: http://www.ghmc.gov.in/Disaster.aspx (accessed July 2018).

HMDA 2019 Master Plan of the City. Available from: hmdaprojects.hmda.gov.in/masterplan (accessed September 2019).

Hosseini, R. H., Golian, S. & Yazdi, J. 2020 Evaluation of data-driven models to downscale rainfall parameters from global climate models outputs: the case study of Latyan watershed. *Journal of Water and Climate Change* **11** (1), 200–216.

Islam, A. R. M. T., 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** (3), 101075.

Jenks, G. F. 1967 The data model concept in statistical mapping. *International Yearbook of Cartography* **7**, 186–190.

Julien, Y., Sobrino, J. A. & Verhoef, W. 2006 Changes in land surface temperatures and NDVI values over Europe between 1982 and 1999. *Remote Sensing of Environment* **103** (1), 43–55.

Kaur, R., Schaye, C., Thompson, K., Yee, D. C., Zilz, R., Sreenivas, R. S. & Sowers, R. B. 2021 Machine learning and price-based load scheduling for an optimal IoT control in the smart and frugal home. *Energy and AI* **3**, 100042.

Kim, S., Seo, Y., Rezaie-Balf, M., Kisi, O., Ghorbani, M. A. & Singh, V. P. 2019 Evaluation of daily solar radiation flux using soft computing approaches based on different meteorological information: peninsula vs continent. *Theoretical and Applied Climatology* **137** (1–2), 693–712.

López, A. S. V. & Rodriguez, C. A. M. 2020 Flash flood forecasting in São Paulo using a binary logistic regression model. *Atmosphere* **11** (5), 473.

Modaresi, F., Araghinejad, S. & Ebrahimi, K. 2018 A comparative assessment of artificial neural network, generalized regression neural network, least-square support vector regression, and K-nearest neighbor regression for monthly streamflow forecasting in linear and nonlinear conditions. *Water Resources Management* **32** (1), 243–258.

Mosavi, A., Hosseini, F. S., Choubin, B., Goodarzi, M., Dineva, A. A. & Sardooi, E. R. 2021 Ensemble boosting and bagging based machine learning models for groundwater potential prediction. *Water Resources Management* **35** (1), 23–37.

Nayan, N. K., Das, A., Mukerji, A., Mazumder, T. & Bera, S. 2020 Spatio-temporal dynamics of water resources of Hyderabad Metropolitan Area and its relationship with urbanization. *Land Use Policy* **99**, 105010.

Ni, L., Wang, D., Wu, J., Wang, Y., Tao, Y., Zhang, J. & Liu, J. 2020 Streamflow forecasting using extreme gradient boosting model coupled with Gaussian mixture model. *Journal of Hydrology* **586** (124), 901.

Oliveira, L. A. B. D. & Carneiro, C. D. C. 2021 Synthetic geochemical well logs generation using ensemble machine learning techniques for the Brazilian pre-salt reservoirs. *Journal of Petroleum Science and Engineering* **196**, 108080.

Pirnia, A., Darabi, H., Choubin, B., Omidvar, E., Onyutha, C. & Haghighi, A. T. 2019 Contribution of climatic variability and human activities to stream flow changes in the Haraz River basin, northern Iran. *Journal of Hydro-Environment Research* **25**, 12–24.

Pradhan, B. 2010 Flood susceptible mapping and risk area delineation using logistic regression, GIS and remote sensing. *Journal of Spatial Hydrology* **9** (2), 1–18.

Raghavendra, N. S. & Deka, P. C. 2014 Support vector machine applications in the field of hydrology: a review. *Applied Soft Computing* **19**, 372–386.

Rahmati, O., Pourghasemi, H. R. & Melesse, A. M. 2016 Application of GIS-based data driven random forest and maximum entropy models for groundwater potential mapping: a case study at Mehran Region, Iran. *Catena* **137**, 360–372.

Ren, M., Pang, B., Xu, Z., Yue, J. & Zhang, R. 2019 Downscaling of daily extreme temperatures in the Yarlung Zangbo River Basin using machine learning techniques. *Theoretical and Applied Climatology* **136** (3–4), 1275–1288.

Sankaranarayanan, S., Prabhakar, M., Satish, S., Jain, P., Ramprasad, A. & Krishnan, A. 2020 Flood prediction based on weather parameters using deep learning. *Journal of Water and Climate Change* **11** (4), 1766–1783.

Sannigrahi, S., Bhatt, S., Rahmat, S., Uniyal, B., Banerjee, S., Chakraborti, S. & Bhatt, A. 2018 Analyzing the role of biophysical compositions in minimizing urban land surface temperature and urban heating. *Urban Climate* **24**, 803–819.

Sarzaeim, P., Bozorg-Haddad, O., Bozorgi, A. & Loáiciga, H. A. 2017 Runoff projection under climate change conditions with data-mining methods. *Journal of Irrigation and Drainage Engineering* **143** (8), 04017026.

Shahabi, H., Shirzadi, A., Ghaderi, K., Omidvar, E., Al-Ansari, N., Clague, J. J., Geertsema, M., Khosravi, K., Amini, A., Bahrami, S., Rahmati, O., Habibi, K., Mohammadi, A., Nguyen, H., Melesse, A. M., Ahmad, B. B. & Ahmad, A. 2020 Flood detection and susceptibility mapping using sentinel-1 remote sensing data and a machine learning approach: hybrid intelligence of bagging ensemble based on k-nearest neighbor classifier. *Remote Sensing* **12** (2), 266.

Swathi, V. 2020 *Analysing the Linkages Between Urban Floods, Climate Change and Land Use*. PhD Thesis, BITS Pilani Hyderabad Campus.

Tehrany, M. S., Pradhan, B., Mansor, S. & Ahmad, N. 2015 Flood susceptibility assessment using GIS-based support vector machine model with different kernel types. *Catena* **125**, 91–101.

USGS 2016 ASTER Data. Available from: http://earthexplorer.usgs.gov/ (accessed January 2016).

Vemula, S., Raju, K. S., Veena, S. S. & Kumar, A. S. 2019 Urban floods in Hyderabad, India, under present and future rainfall scenarios: a case study. *Naturals Hazards* **95**, 637–655.

Wang, Z., Lai, C., Chen, X., Yang, B., Zhao, S. & Bai, X. 2015 Flood hazard risk assessment model based on random forest. *Journal of Hydrology* **527**, 1130–1141.

Wu, L., Peng, Y., Fan, J. & Wang, Y. 2019 Machine learning models for the estimation of monthly mean daily reference evapotranspiration based on cross-station and synthetic data. *Hydrology Research* **50** (6), 1730–1750.

Yin, J., Ye, M., Yin, Z. & Xu, S. 2015 A review of advances in urban flood risk analysis over China. *Stochastic Environmental Research and Risk Assessment* **29** (3), 1063–1070.

Zhou, J., Jia, L. & Menenti, M. 2015 Reconstruction of global MODIS NDVI time series: performance of Harmonic ANalysis of Time Series (HANTS). *Remote Sensing of Environment* **163**, 217–228.

First received 16 February 2021; accepted in revised form 30 March 2021. Available online 12 April 2021