An approach for Flood Severity assessing Flood Impact Prediction using Hybrid Machine Learning Models

Phulre Ajay Kumar[1]*, Khekare Ganesh[2], Ameta Gaurav Kumar[3], Jain Ankur[1] and Joshi Suneet[1]
1. School of Computing Science Engineering and Artificial Intelligence, VIT Bhopal University, Sehore M.P., INDIA
2. Vellore Institute of Technology, Vellore, INDIA
3. Parul University, Vadodara, INDIA
*ajaykumarphulre@vitbhopal.ac.in

Abstract

One major natural disaster that happens all around the world as a result of natural forces is flooding. They have resulted in animal losses, significant property damage and even fatalities. It is crucial to have a flood risk prediction system that is both accurate and effective. Early warning system helps to minimize any harm. It is crucial for protecting people and property by sending out timely alerts. This study employs and evaluates four different machine learning algorithms: Decision Tree Classifier, Random Forest Classifier, Logistic Regression and K-Nearest Neighbors (KNN) Classifier. Additionally, it includes a comprehensive analysis aimed at assessing flood susceptibility across the targeted region.

Evaluation of these models' predictive power for flood susceptibility is the goal. The algorithms' classification accuracies for the provided datasets are 62%, 87%, 83% and 83% respectively. This study underscores the role of machine learning approaches in disaster management. It specifically delves into scholarly work on hazard prediction, disaster detection, early warning systems, monitoring, risk and vulnerability assessment, damage appraisal, post-disaster recovery and pertinent case studies.

Keywords: Machine Learning Algorithms, Natural Disaster Management, Flood Risk Assessment, Classification, Data Analytics, Emergency Management.

Introduction

Floods are one of the most devastating natural calamities which have tremendously impacted agriculture as well as infrastructure, human lives and the socio-economic fabric. Therefore, the Governments are put under more and more pressure to provide reliable and accurate flood risk maps and to support long–term flood risk management plans focusing on protection, prevention and preparation. Flood prediction models are essential in ensuring the ability to cope with extreme weather events and assessing the dangers of flooding. They are critical in the development of efficient evacuation plans, policy recommendations and full scale plans for water resource management. The first case study considers an empirical dataset to classify flood severity via a set of machine learning techniques. In this dataset, there is a range of elements that are necessary in ascertaining the severity of floods such as the length, magnitude and the water level of each occurrence. Levels of flooding are as follows: The regular flood is class one, an unusual flood is class two and an extremely dangerous flood level is class three. The datasets are organized as a time series data for this purpose. In this way, it is expected that the study will make flood severity classification more accurate so as to come up with better flood management and prediction methods.

Machine learning (ML) has become popular among the hydrologists as it has a lot to offer. To develop more accurate and efficient models that they can use in flood prediction, researchers keep developing new machine learning approaches and incorporating pre-existing ones. The key objective of this research is thus to study the State-of-the-Art machine learning technologies employed for flood prediction and to find out the models that generate the most precise and reliable results. This study is a comprehensive research on the various ML approaches ML models reviewed on the basis of robustness, accuracy, efficiency and speed[21]. With critical evaluation and discussion, comparative analysis of machine learninig models will lead to the deep understanding of these sets of methodologies[10].

This research carries out the critical evaluation of advanced techniques (artificial intelligence, machine learning, Internet of things (IoT), cloud computing and robotics) for forecasting flash floods[3]. It provides the best methods in forecasting both short and long-term flood events. The evaluation made by the study shows significant improvements in flood prediction like the use of hybrid models, data decomposition techniques, algorithm ensembles techniques and model optimization. This review can be used as a useful source of information for climate scientists and hydrologists, in terms of choosing the most appropriate machine learning strategies for particular application of predictions[7].

The susceptibility of the plantation industry is demonstrated by research on the economic impacts of floods in some parts of Kerala. Moreover, the data-driven models have for long helped the flood modeling by enhancing the numerical and physical models. Precise intensity assessment and classification of natural catastrophes become possible with such approaches, for example, multispectral picture analysis based on multi-layer deep convolutional neural networks (CNNs). These procedures are increasingly in vogue

**** Author for Correspondence***

https://doi.org/10.25303/188da0108

because they are able to predict floods better by using climatic indices and the data of hydro-meteorological. Certain benefits of these strategies are also illustrated by another perspective on the problem of catastrophe risk reduction.

Some of the most popular statistical methods used in flood frequency analysis (FFA) include the use of Autoregressive Integrated Moving Average (ARIMA), Multiple Linear Regression (MLR) and Autoregressive Moving Average (ARMA). These models are the base of the classical statistical method of flood forecasting and they are an essential component of the structure of hydrologic prediction frameworks. Nevertheless, limitations of these conventional and physically based models have exponentially increased the need for advanced data-driven methods, mainly, ML models. This trend can be observed in the recent studies devoted to the flood risk assessment based on algorithms such as Random Forest for a better predictive effectiveness[11].

The ML models have received growing attention due to their ability to accurately model the nonlinearity of flood dynamics experienced with only the historical data. This is demonstrated by the applicability of XGBoost algorithm in landslide susceptibility mapping in the upper basin of Ataturk Dam in Turkey. These models are ranked as very promising for flood prediction based on their fast processing and light weightiness in requirement of input[6]. Machine learning is one subfield of artificial intelligence (AI) that is able to extract patterns and trends, where it can usually perform better than conventional physical models which offer faster training, validation, testing as well as evaluation processes with lower computational needs.

Some other studies concerning Random Forest algorithms explain how this quality allows ML-based models to achieve outstanding results in the flood catastrophe risk assessment[13]. The impact and the key factors of digital transformation in disaster management have been discussed in the context of UK's national experience in dealing with disasters[21]. Advances in the field of machine learning during the last two decades have demonstrated its usefulness in flood forecasting. As compared to the tradition techniques, it often wins in efficacy and precision. Issues of interest, leading research directions and upcoming projects have been determined by those studies which focus on computational intelligence as an integral part of enormous scale flood control[4]. Computational intelligence is seen to be a measure to enhance resilience to major flood occurrences and to strengthen disaster management plans.

The efficiency of ML and physical prediction models had recently been studied[2] where it has been revealed that the ML models deliver higher precision. In addition, various studies have indicated satisfactory ML techniques' implementation in QPF for different lead-time intervals. Moreover, a promise has been made in the use of a fuzzy expert system for automatic selection of wavelet shrinkage processes to reduce noises in this kind of forecasts[8,12].

ML models are always able to produce better predictions compared to traditional statistical models[5]. Ortiz-García and colleagues[13] showed how such complicated hydrological processes such as floods can be effectively modelled via machine learning approaches. Prediction of the growth of sugarcane with the help of artificial neural networks and extreme learning machines is based on the climatic conditions[21]. Support vector regression is for regional flood frequency analysis under past and future climatic condition[14]. Wavelet-linear genetic programming model as a new approach to simulating monthly stream flows has been presented by us[19]. Broad general uses of machine learning in flood prediction are also not well-investigated in literature which present a gap.

Dataset Description

The worst floods that had happened in the southern Indian State of Kerala in over a century were registered in August, 2018, when excessively torrential rains during monsoon season caused the catastrophic flooding. Apart from causing close to a million of the citizens to evacuate, the disaster killed over 483 people and left another 140 people as missing. All the fourteen districts in Kerala were on a red alert. The floods and other disasters affected at least one sixth of the population in Kerala. Being a "calamity of severe nature", the incident was given level 3 calamity by the Indian Government.

In the figure 1, the rain fall index for July, August, September and October has been represented. Some of the data that make up the dataset are: Data of 2018 Kerala floods such as rainfall, district-specific alerts/advisories and people deaths. The titles of the columns give an indication of the content in clear terms. It is very important to analyze the changes in raining fall index in the period of rainy season, especially in August, September and October for forecasting floods. Because as it happens during most of these months, the monsoon season reaches its peak in many parts of the world. Monitoring rainfall patterns will help in issuing early flood warnings. August is usually the peak of monsoon with heavy rainfall. The dataset tells us how continuous heavy showers tend to significantly increase the rainfall index at this period of time. From September, rainfall patterns can change, but there is a possibility for significant rain events, but with the decrease in general intensity as compared to August.

Material and Methods

Flood prediction often relies on real-time observations made in a variety of return times obtained from several rain gauges or other sensing equipment. Historically, this information has come from rainfall and water level readings and these are normally measured using advanced remote sensing tools such as satellites, multisensor setups, radar or a ground rain gauge.

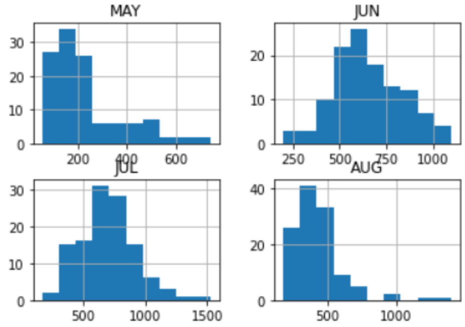


Fig. 1: The rainfall index varying during rainy season

Remote sensing is a common method of getting real-time data with a higher resolution. In fact, readings obtained from radar studies are often found to be more reliable than the readings from standard rain gauges. Using radar rainfall therefore, prediction models have been more effective. Whatever, whether it is radar or ground-based information, historical records remain necessary for constructing predictive models.

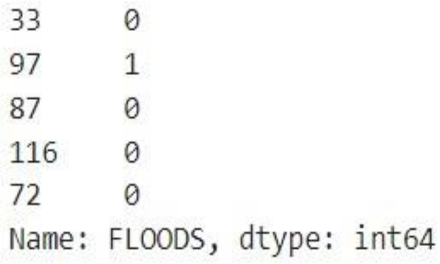
In machine learning based flood prediction model, the data records i.e. hourly to monthly are split into train, validation and test sets to robust model development. Many well-known techniques of the ML have been widely used in this domain, namely Artificial Neural Networks (ANNs), Neuro-Fuzzy Systems, the Adaptive Neuro-Fuzzy Inference Systems (ANFIS), Support Vector Machines (SVM), Wavelet Neural Networks (WNN) and Multilayer Perceptrons. The methodology is a foundation for the development of effective strategies and practical recommendations that are targeted at enhancing disaster management initiatives.

This paradigm would help to analyze the already existing systems and areas for improvements to facilitate effective planning, preparedness as well as a response to cases of disasters. It enables a thorough evaluation of risks, vulnerabilities and possible challenges so that the strategies of managing disasters are customized to certain eventualities or demands. Such methodology may also be useful in tuning up coordination between various agencies, resource utilization and enhancing communication during disaster events. It emphasizes proactiveness with a consideration of both prevention and mitigation so as to reduce the effects of anticipated disasters. Also, it enables to develop flexible and scalable plans that are to be altered according to the gravity of the situation.

Through the cooperation between governmental bodies, non-governmental organizations and local communities, such an approach will create a more cohesive and efficient system of the disaster response. Finally, during emergencies, such an approach helps to save lives and property, reduces the levels of risks and enhances skills in handling the disaster. The application of real-time accumulative data retrieved from various rainfall gauges and sensing devices for diverse returns is a norm for predicting floods. These datasets are usually obtained from ground observation rain-gauge and up-to-date remote sensing instruments such as satellites, radar and multi-sensor systems.

Specifically, remote sensing is mostly preferred for gathering high-resolution real-time data. In comparison with other methods of the rainfall observations, radar's observations are much more consistent and accurate in the measurement of the rainfall than those of the traditional rain gauges. Therefore, using radar rainfall data in predictive models has been found to increase their accuracy. Historical data sources, be it radar systems or ground-level gauges, are vital in the development of such models.

Divide the dataset into 2 sets:train and test in ratio (4:1)



```
33     0
97     1
87     0
116    0
72     0
Name: FLOODS, dtype: int64
```

Fig. 2: Train set of data

Results and Discussion

Historical datasets (providing hourly, daily, or monthly values) are typically divided into training, validation, verification and testing subsets to develop productive machine learning model. The processes and techniques of flood prediction using machine learning are well documented. Types of techniques that are commonly used in this area of work include: Artificial Neural Networks (ANNs), Neuro-Fuzzy Systems, Adaptive Neuro-Fuzzy Inference Systems (ANFIS), Support Vector Machines (SVM), Wavelet Neural Networks (WNN) and Multilayer Perceptrons (MLP). Some of the machine learning methods that are used here, include Random Forest Classification, Decision Tree Classification, Logistic Regression and K-Nearest Neighbors (KNN) Classifier. The bottom line of building the machine learning model can be seen from fig. 3.

https://doi.org/10.25303/188da0108

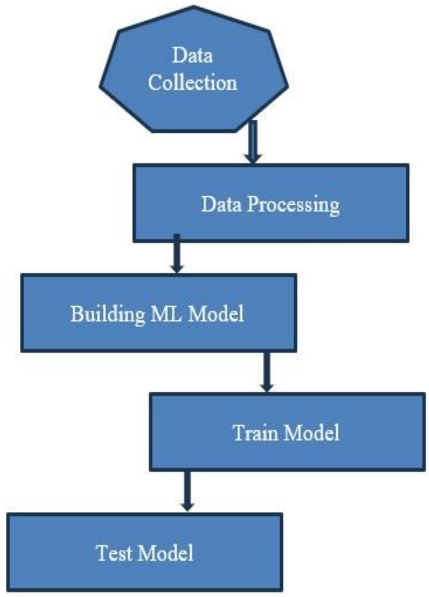


Fig. 3: Basic flow for building the machine learning (ML) model

KNN-Classifier and Logistic regression model: The K-Nearest Neighbor (KNN) algorithm is one of the most popular supervised machine learning algorithms that can be used for both kinds of tasks, classification as well as regression. It works by keeping all instances of training in a multidimensional feature space in which each point is an instance described by different attributes. When a new completely unobserved example enters the processing, the algorithm finds the k most similar examples (ref. to "neighbors") based on the stored training data by comparing it to the incoming example. In case of real-valued attributes, Euclidean distance is usually used for computing similarity. In regression tasks, KNN determines the outcome by averaging the value of these k nearest neighbours.

Logistic regression is used in examining the relationships between one or more than one predictor variables (whether continuous or categorical) and a categorical outcome variable, whether nominal or ordinal. It is a basic classification technique that allows measuring the relationship existing between a predictor variable (x) and a response variable (y), which can either be dichotomous or the response variable may be polycotomous. If the response variable is binary logistic regression, it is divided into two categories, usually called "success" and "failure," with y = 1 for success and y = 0 for failure. The model computes the probability of the response from Bernoulli distribution of outcome variable for each observation.

If y = 0, then f(y) = 1 - π and if y = 1, then f(y) = π. From this, the logistic regression function can be expressed as follows:

$$f(z) = \frac{e^{z}}{1+e^{z}}; z = a + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

where n represents the number of predictor variables. The value of f(z) ranges between 0 and 1 for any given z, so z might be anywhere between -∞ and ∞. In essence, the logistic regression model shows the likelihood that a flood will occur. The following is the definition of the logistic regression model:

$$\pi(x) = \frac{e^{a+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n}}{1+e^{a+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n}}$$

Since the function π(x) is nonlinear, a logit transformation is required to change it into a linear form. This makes it possible to comprehend the relationship between the predictor variable and the response variable, y more clearly.

$$= a + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

Logistic regression analysis of this study made use of data from flood events in the form of dependent variables. This approach allows building the flood danger map that reflects the local peculiarities accurately. This study assigns weights that are based on observed flood data and not the traditional weight values which are predetermined. To make sure that the coefficients of the equation reflect real flood events, 70% of the total flood event data is randomly chosen as material for the model.

Decision tree classification and Random Forest algorithm Decision tree are popular supervised learning techniques that can be used in questions of regression problems or classification problems. It is constructed as a tree-like structure, where the leaf node contains expected outcome while the branches identify decision-making pathways and each internal node corresponds to features of a given dataset.

This structure comprises of Nodes which are the main node types where the nodes make decisions on the basis of the dataset and include several branches and the other nodes which retain the final conclusion and have no further branches. The choices in a decision tree are based on the analysis of certain properties of any dataset. This algorithm creates a tree-like form of arrangement that presents the possible outcomes as regards to certain parameters; it is done from a root node building up into branches.

The tree is built with the aid of the classification and regression tree (CART) algorithm, which divides the data into the smaller subtrees considering the binary decisions. Decision trees are mainly employed because they possess an easily explainable tree form and due to their ability to emulate the human decision-making. Starting from the root node, the method examines the dataset properties in order to determine an optimal way. It then compares until it gets to a leaf node where the final choice is made. Random Forest is a well-known machine learning method that falls under the category of supervised learning approaches. It can handle tasks involving both regression and classification and is flexible.

The approach is based on ensemble learning, which combines several classifiers to address difficult issues and increase model efficacy. As the name suggests, Random Forest is a collection of decision trees built from different data subsets. The model improves forecast accuracy by averaging these decision tree results. The steps involved in the Random Forest algorithm's operation are depicted in figure 4.

Performance Evolution: It describes the experimental setup, covering the design of the test environment, the configuration of the models and the specific models used in the experiments. Additionally, it explains the performance evaluation methods applied to analyze the machine learning algorithms on the flood forecasting datasets. Accuracy is 0.8333333333333334. For a given collection of data points, the graph shows a comparison between the actual and expected possibilities of flooding. When assessing the model's accuracy and error performance, this graphic is useful.

Logistic regression: The aforementioned graph shows a comparison of each data point's actual flood probability (shown by a dotted line) and expected flood probabilities (shown by a regular line). The two sets of values may be easily distinguished from one another thanks to the markers.

Accuracy score: 83.333333
Recall score: 76.923077
Roc score: 83.916084

Decision tree classification: The graphic contrasts the actual and expected probabilities of flooding at each observation site. Crosses ('x') indicate the actual values whereas circles ('o') indicate the expected values. It is simple to evaluate how well the forecasts match the actual results.

Accuracy score:62.500000
Recall score:46.153846
Roc score:63.986014

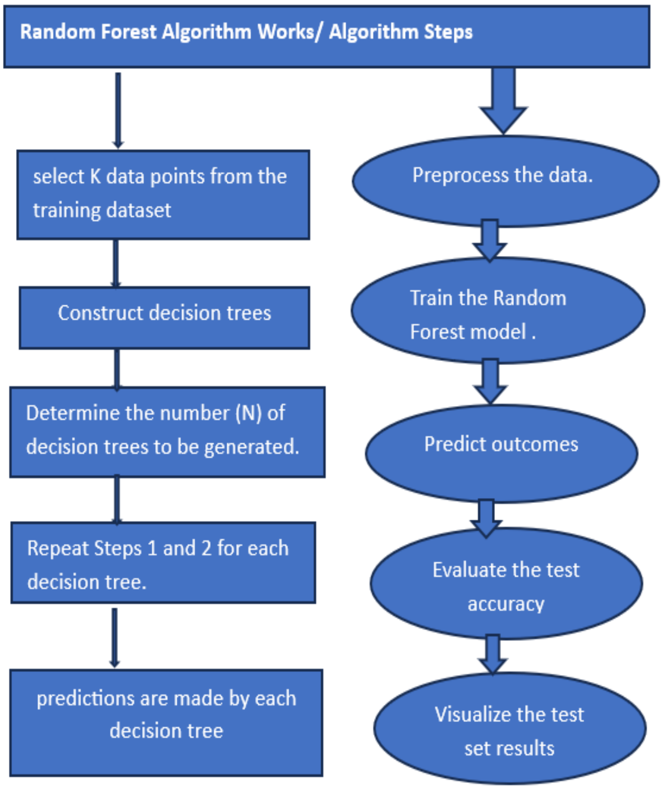

Random Forest Algorithm Works/ Algorithm Steps
select K data points from the training dataset
Construct decision trees
Determine the number (N) of decision trees to be generated.
Repeat Steps 1 and 2 for each decision tree.
predictions are made by each decision tree
Preprocess the data.
Train the Random Forest model .
Predict outcomes
Evaluate the test accuracy
Visualize the test set results

Fig. 4: Random forest algorithm working and process step

Table 1
Predict chances of flood KNN Classifier

Floods, d type: int 64	Actual chances of flood	Predicted chances of flood
80	1	1
21	1	1
68	0	0
7	0	0
56	1	1
62	0	0
64	0	0
0	1	1
37	0	0
43	0	0
71	0	1
55	0	0
78	0	0
105	1	1
14	1	1
32	1	1
74	1	1
94	0	1
86	0	0
57	1	1
58	1	1
95	0	0
39	1	0

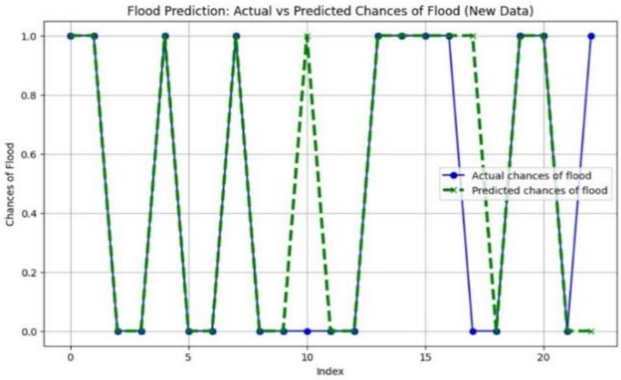


Fig. 5: KNN predicted values Vs actual values Accuracy KNN

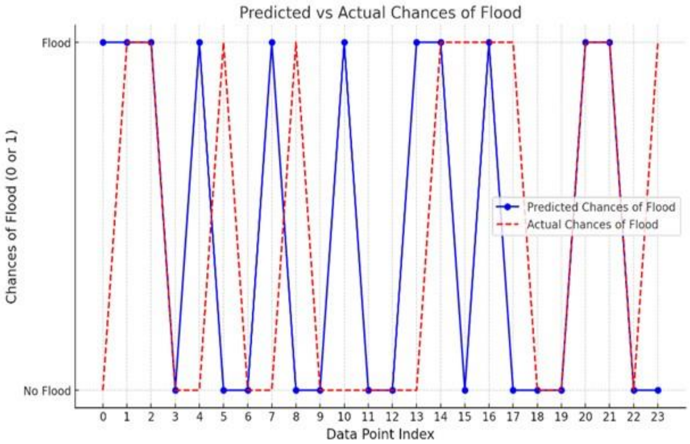


Fig. 6: Logistic regression predicted values Vs actual values

https://doi.org/10.25303/188da0108

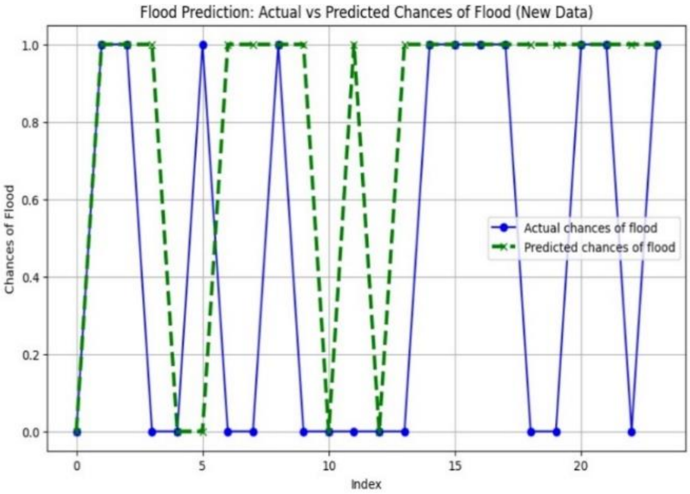


Fig. 7: Decision tree classification Logistic regression

Random Forest Classification: The majority of cases are accurately classified by the model, which has an accuracy of 87.5%. It misses some affirmative cases, as indicated by its recall score of 76.92%. The model's great discriminating strength and ability to differentiate across classes are demonstrated by its ROC AUC score of 88.46%.

Random Forest Classifier (max_depth=3, random_state=0)
Accuracy score: 87.500000
Recall score: 76.923077
Roc score: 88.46153

Conclusion

The methodology offered in this project produced promising outcomes in terms of predicting floods based on scarce sensor readings. It can be demonstrated with the help of this study that the logistic regression machine learning model had highest accuracy as compared to other models. With the use of minimal data set, the model could identify relevant predictors and conditions for flood scenarios. The purpose of this study is developing the predictive model for estimating flood risks in Kerala. Results showed that there was a slight improvement in logistic regression when it is applied on the standard data and models' predictive accuracy was compared.

A dynamic system that can take account of variations in time should be established because floods are events that take a short time and immediate preventive responses are needed. Floods do not only lead to massive damages of properties but can also lead to enormous loss of human life and therefore prompt action is necessary.

Recommendations

The model can be augmented with an addition of a misclassification cost function that has not been covered in this study. This aspect may assist in enhancing resource distribution in flood management strategies.

- Future research should be oriented towards an optimization of the cost function in order to minimize the risks of misclassification of a disaster event, with some severe social and economic consequences.
- It is important to examine its influence more carefully since it is possible to cause devastating effects by making mistakes in predicting floods to the people who suffer.
- More studies concerning weights employed in the computation of the flood risk index are required to determine the best values capable of increasing the predictive ability of the model.
- Competent research on the adjustment of these weights could be beneficial towards the full optimization of the performance of the model so that the early warning systems for flood events are promoted.

References

1. Aamir M., Ali T., Irfan M., Shaf A., Azam M.Z., Glowacz A., Brumercik F., Glowacz W., Alqhtani S. and Rahman S., Natural Disasters Intensity Analysis and Classification Based on Multispectral Images Using Multi-Layered Deep Convolutional Neural Network, *Sensors*, **21(8)**, 2648 **(2021)**

2. Alexander D. and Davis I., Disaster risk reduction: An alternative viewpoint, *International Journal of Disaster Risk Reduction*, **2**, 1–5 **(2022)**

3. AlHinai Y., Disaster management digitally transformed: Exploring the impact and key determinants from the UK national disaster management experience, *International Journal of Disaster Risk Reduction*, **51**, 101851 **(2020)**

4. Al-Rawas G., Nikoo M.R., Al-Wardy M. and Etri T., A critical review of emerging technologies for flash flood prediction: Examining artificial intelligence, machine learning, Internet of Things, cloud computing and robotics techniques, *Water*, **16(14)**, 2069 **(2024)**

5. Banna M., Taher K.A., Kaiser S., Mahmud M., Rahman S., Hosen S.A. and Cho H., Application of Artificial Intelligence in Predicting Earthquakes: State-of-the-Art and Future Challenges, *IEEE Access*, **8**, 192880–192923 **(2020)**

6. Cai H., Research on flood disaster risk assessment based on Random Forest algorithm, In IEEE 2nd International Conference on Data Science and Computer Application (ICDSCA) **(2022)**

7. Can R., Kocaman S. and Gokceoglu C., A comprehensive assessment of XGBoost algorithm for landslide susceptibility mapping in the upper basin of Ataturk Dam, Turkey, *Applied Sciences*, **11(11)**, 4993 **(2021)**

8. Dineva A., Várkonyi-Kóczy A.R. and Tar J.K., Fuzzy expert system for automatic wavelet shrinkage procedure selection for noise suppression, In 2014 IEEE 18th International Conference on Intelligent Engineering Systems (INES), Tihany, Hungary, 163–168 **(2014)**

9. Fotovatikhah F., Herrera M., Shamshirband S., Chau K.W., Faizollahzadeh Ardabili S. and Piran M.J., Survey of computational intelligence as basis to big flood management: Challenges, research directions and future work, *Eng. Appl. Comput. Fluid Mech*., **12**, 411–437 **(2018)**

10. Gizaw M.S. and Gan T.Y., Regional flood frequency analysis using support vector regression under historical and future climate, *J. Hydrol*., **538**, 387–398 **(2016)**

11. Kim S., Matsumi Y., Pan S. and Mase H., A real-time forecast model using artificial neural network for after-runner storm surges on the Tottori Coast, Japan, *Ocean Eng*., **122**, 44–53 **(2016)**

12. Mosavi A., Rabczuk T. and Varkonyi-Koczy A.R., Reviewing the novel machine learning tools for materials design, In Recent Advances in Technology Research and Education, Springer, Cham, Switzerland, 50–58 **(2017)**

13. Ortiz-García E., Salcedo-Sanz S. and Casanova-Mateo C., Accurate precipitation prediction with support vector classifiers: A study including novel predictive variables and observational data, *Atmos. Res*., **139**, 128–136 **(2014)**

14. Phulre A.K., Pagare S. and Chakrawati A., Automated Framework for Web Content Security Through Content Management System, 10th International Conference on Emerging Trends in Engineering and Technology - Signal and Information Processing (ICETET-SIP-22), Nagpur, India, 1–4 **(2022)**

15. Phulre A.K., Kamble M. and Phulre S., Content Management Systems hacking probabilities for Admin Access with Google Dorking and database code injection for web content security, 2nd International Conference on Data, Engineering and Applications (IDEA), Bhopal, India, 1–5 **(2020)**

16. Phulre A.K., Jain S. and. Jain G, Evaluating Security enhancement through Machine Learning Approaches for Anomaly Based Intrusion Detection Systems, 2024 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS), Bhopal, India, 1–5 **(2024)**

17. Phulre A.K. and Kamble M., Study and Analysis of Web Content Security Through Content Management Systems, *International Journal of Emerging Technology and Advanced Engineering*, **9(10)**, 99–103 **(2019)**

18. Phulre A.K., Verma M., Mathur J.P.S. and Jain S., Approach on Machine Learning Techniques for Anomaly-Based Web Intrusion Detection Systems: Using CICIDS2017 Dataset, In MAiTRI 2023, Lecture Notes in Networks and Systems, Springer, Singapore **(2024)**

19. Ravansalar M., Rajaee T. and Kisi O., Wavelet-linear genetic programming: A new approach for modeling monthly streamflow, *J. Hydrol*., **549**, 461–475 **(2017)**

20. Sabu G. and Manoj G., The economic impact of floods on the plantation sector: a study of selected districts in Kerala, *Disaster Advances*, **16(3)**, 13–22 **(2023)**

21. Taherei Ghazvinei P., Hassanpour Darvishi H., Mosavi A., Yusof K.B.W., Alizamir M., Shamshirband S. and Chau K.W., Sugarcane growth prediction based on meteorological parameters using extreme learning machine and artificial neural network, *Eng. Appl. Comput. Fluid Mech.*, **12**, 738–749 **(2018)**.

(Received 10th March 2025, accepted 08th May 2025)