e-ISSN:2582-7219

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

Volume 7, Issue 13, April 2024

INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

Impact Factor: 7.521

☎ 6381 907 438 6381 907 438 ijmrset@gmail.com @ www.ijmrset.com

Machine Learning Approaches for Flood and Cyclone Prediction

S. Kanmani[1], R. Lakshmi Priya[2], M. Baanupriya[3], S. Selvy[4], S. Rohan[5]

Department of Information Technology, Puducherry Technological University, Puducherry, India[1,2,3,4,5]

ABSTRACT: Floods and cyclones, two of nature's most destructive phenomena, pose significant threats to communities and infrastructure worldwide. Floods result in immense damage to land, buildings, and human fatalities, while cyclones bring about devastating winds, storm surges, and heavy rainfall. In our work, we leverage machine learning techniques to predict these disasters, focusing on MLP, Extra-Tree Classifier and Catboost for flood prediction and parallel LSTM and CNN models for cyclone prediction. By integrating these models into a user-friendly interface, our aim is to advance predictive modeling techniques and minimize the impact of floods and cyclones on vulnerable areas. Our research contributes to enhancing disaster preparedness and decision-making processes for policymakers, emergency responders, and community leaders. Through the implementation of proactive measures, we strive to mitigate the devastating effects of floods and cyclones on society. Our findings provide valuable insights into the application of machine learning for disaster prediction and management, ultimately leading to more resilient communities and infrastructure in the face of natural disasters.

KEYWORDS: Flood, Cyclone, Machine Learning models.

I. INTRODUCTION

The prediction of natural disasters, such as floods and cyclones, has become increasingly important in recent years due to their devastating impact on communities and infrastructure. Just as Bitcoin's price fluctuations garner attention in the digital currency realm, the unpredictability of these disasters captures the interest of scientists, policymakers, and the public alike. A myriad of factors, including climatic conditions, environmental changes, and human activities, influence the occurrence and severity of floods and cyclones. However, predicting these events with precision remains challenging due to their complex and dynamic nature. Various techniques, such as machine learning algorithms and predictive modelling, have been employed to forecast flood and cyclone occurrences. Yet, the inherent volatility and decentralization of these disasters pose significant obstacles to accurate predictions. In this context, our work seeks to leverage advanced machine learning approaches to enhance the prediction accuracy of floods and cyclones, thereby aiding in disaster preparedness and mitigation efforts.

II. RELATED WORK

A literature review, sometimes known as a survey, looks at the body of knowledge regarding a certain subject. It is crucial for giving readers a thorough grasp of the topic, pointing out areas in need of further investigation, and placing recent findings in the context of earlier research. This section offers a brief summary of the literature on bitcoin price prediction that is currently available, looking at the many approaches used in these fields.

The paper [1] focuses on flood prediction using three machine learning models: MLP classifier, CatBoost classifier, and Extra-Tree classifier, with rainfall data as the training dataset. CatBoost outperformed the other models, achieving an accuracy of 98.34%, compared to MLP (94.5%) and Extra-Tree (97.9%). The comparative analysis emphasizes CatBoost's consistent superiority in accuracy scores, while the Extra-Tree classifier showed fewer false positives. Evaluation of performance metrics such as precision, ROC, recall, and accuracy consistently favored CatBoost, highlighting its efficacy in flood prediction based on rainfall data.

The paper [2] presents a flood prediction system combining Machine Learning (ML) classifiers and GIS techniques for urban management and resilience planning. Utilizing a Random Forest model, the study achieves high performance with a Matthew's Correlation Coefficient of 0.77 and an Accuracy of 0.96. The integration of GIS identifies floodprone areas based on historical data, creating a comprehensive flood risk index by combining ML scores and GIS results. The research highlights the significance of rainfall as the most effective predictor, advocating for enhanced predictive power through feature engineering. Evaluation metrics include Accuracy, AUC, Recall, F1, and MCC, showcasing the system's efficacy in urban flood prediction and management.

The paper [3] focuses on predicting flash floods using Indian district rainfall data and machine learning algorithms, including Linear Regression, K-Nearest Neighbor, Support Vector Machine, and Multilayer Perceptron (MLP). The MLP model stands out with an impressive accuracy of 97.40%, offering a valuable tool for climate scientists to predict floods during heavy downpours. The dataset is divided into training and testing sets, and the model's performance is evaluated using a confusion matrix. The proposed MLP-based model not only demonstrates high accuracy but also provides a straightforward and efficient approach to flash flood prediction based on rainfall data, highlighting its practical utility in the field.

The paper [4] focuses on tropical cyclone prediction in the Indian coastal region, employing diverse deep learning networks, including MLP, LSTM, GRU, RNN, BI-LSTM, and CNN. After rigorous evaluation, CNN emerges as the most effective model, prompting further analysis. The model's hyperparameters undergo optimization using a genetic algorithm. Notably, the conventional fully connected layer in the CNN model is replaced with various machine learning classifiers, such as Decision Tree, K-Nearest Neighbor, Logistic Regression, Naive Bayes, Random Forest, SVM, and XGBoost. The unique adaptation of the C4.5 Decision Tree algorithm within the CNN model enhances prediction accuracy. The proposed model is tested on five cyclones, validated against the Saffir-Simpson scale, and exhibits superior performance in time complexity, accuracy, precision, and recall compared to both traditional machine learning and deep learning classifiers. This comprehensive approach offers a promising and efficient solution for tropical cyclone prediction in the Indian coastal region.

The paper [5] focuses on flood susceptibility modeling in the Teesta River basin, Bangladesh, utilizing advanced ensemble machine learning models. Introducing two innovative hybrid ensembles, Dagging and Random Subspace (RS), alongside established models like Artificial Neural Network (ANN), Random Forest (RF), and Support Vector Machine (SVM), the study integrates twelve flood-influencing factors and GIS technology. Validation and comparison through statistical measures such as Freidman, Wilcoxon signed-rank, t-paired tests, and Receiver Operating Characteristic Curve (ROC) reveal Dagging as the superior model, followed by RF, ANN, SVM, and RS. The insights gained provide valuable contributions to flood disaster management and the formulation of effective mitigation strategies in the region.

The paper [6] focuses on predicting Climate-Induced Disasters (CID) by establishing a linkage between climate change indices and historical disaster records. Employing a deep learning model trained on flood disaster data from the Canadian Disaster Database in Ontario, the study achieves a notable accuracy of around 96% in predicting flood occurrences. Noteworthy findings include the strong association between flood disasters and precipitation indices, as well as temperature-related features such as daily temperature gradient and the duration of sub-zero minimum temperatures. The innovative approach of integrating historical disaster data, global climate models, and climate change metrics offers a fresh perspective on CID prediction. The ultimate goal is to bolster urban resilience and mitigate CID risks globally, providing valuable insights for effective disaster management strategies.

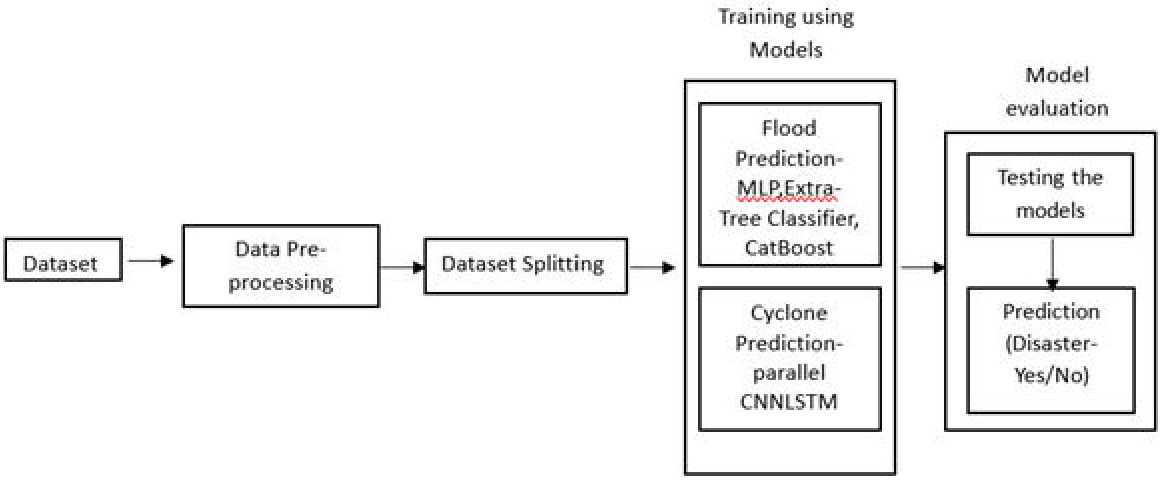
Table 1. Summary of related literature

Authors	Dataset Used	Techniques	Results	Limitation
K Sandhya Rani Kundra, et al.	Rainfall dataset from urban areas in India from 1901 to 2015	MLP,Extra-Tree Classifier and CatBoost	CatBoost model excels with high accuracy in flood prediction	Limited discussion on the impact of external factors..
Marcel Motta, et al.	OpenVC dataset utilized for flood prediction system	ML classifiers, GIS and RandomForest	Combining ML classifiers with GIS enhances urban management efficiency.	Limited to flood prediction, not other natural

				disasters.
Vinothini A,, et al.	https://www.kaggle.com/rajanand/rainfall-in-india.	Linear Regression, K-Nearest Neighbor, SVM, MLP, and Logistic Regression	MLP algorithm with accuracy of 97.40% and various ML approaches enhances accuracy.	Limited execution time details and uncertainties due to climate change impacting rainfall patterns.
P. Varalakshmi, et al.	Meteorological data from MERRA-2 (50 km resolution) and cyclone data from RSMC - New Delhi.	MLP, LSTM, GRU, RNN, BI-LSTM, CNN, C4.5	Modified C4.5 algorithm addresses primary limitations in cyclone prediction.	Limited testing, constraint to only classification rather than more detailed cyclone forecasting.
Abu Reza Md Towfiqul Islama, et al.	OpenVC dataset was utilized to train and enhance the model.	Dagging and Random Subspace coupled with ANN, RF, SVM.	Dagging model outperforms RF, ANN, SVM, RS, and benchmarks, achieving AUC of ROC above 0.80 for all flood susceptibility models.	Limited discussion on the specific challenges faced during flood modeling.
Haggag, M., Siam, A.S., El-Dakhakhni, W. et al	Canadian Disaster Database, created by Public Safety Canada	A deep learning model was developed for spatial-temporal disaster occurrence prediction by linking different climate change indices to historical disaster records.	The deep learning model developed achieved an accuracy of around 96% in predicting flood disasters in Ontario.	May not be directly applicable to other types of climate-induced disasters, Doesn't consider land use, infrastructure, or socioeconomic factors.

III. PROPOSED ALGORITHM

Fig 1 represents the overall flow of the proposed system. The dataset is pre-processed for removing null values. This dataset is then encoded. The pre-processed data is split into training and testing tests. Machine learning models are built and trained using the training dataset. For flood prediction MLP classifier, Extra-tree classifier, Catboost classifier are


Dataset
Data Pre-processing
Dataset Splitting
Training using Models
Flood Prediction-MLP,Extra-Tree Classifier, CatBoost
Cyclone Prediction-parallel CNNLSTM
Model evaluation
Testing the models
Prediction (Disaster-Yes/No)

A. MLP

MLP stands for Multi-layer perceptron. It is a feedforward neural network with fully connected layers with a nonlinear kind of activation function. MLP consist of one input layer, one or more hidden layer and an output layer. MLP learns by adjusting weights during training using backpropagation and gradient descent. It uses an activation function to introduce non-linearity into the model, allowing it to learn complex patterns in the data.

The input layer receives the input data and passes it to the first hidden layer without performing any computations. The hidden layer is where all computations take place. Each neuron in hidden layer receives input from all neurons in the previous layer. Additionally, the neurons also have an associated bias which allows it to adjust its output threshold. The weighted sum of its input is calculated for each neuron in a hidden layer by summing the product of each input and its corresponding weight and adding the bias to it.

$$Weighted\ sum = \sum_{i=1}^{n} (w_i * x_i) + b \quad (1)$$

B. Extra tree classifier

Extremely randomized tree (Extra tree) classifier is an ensemble supervised machine learning technique that trains numerous decision trees and aggregates results from group of decision trees to output its result. It builds the decision trees using random subset of training data and random subset of feature at each split. It randomly selects the splitting value at which to split a feature and create child nodes. This helps in making the tree diversified and uncorrelated.

C. CatBoost

CatBoost or Categorical boosting is an open-source boosting library. It’s a supervised machine learning method. It can handle both numerical and categorical data. It does require encoding to convert categorical data to numerical data. It uses symmetric weighted quantile sketch(SWQS) algorithm to automattically handle the missing values. Catboost uses symmetric trees, which means that all the decision nodes use the same split condition at every depth level.It uses ordered boosting to overcome the problem of overfitting. During training, CatBoost builds an ensemble of decision trees sequentially with each tree learning to correct the errors made by the previous trees.

D. Parallel CNN-LSTM

Parallel CNN-LSTM is a architecture that combines Convolutional Neural Networks (CNNs) and Long Short-term memory (LSTMs) networks in parallel. CNN are expert in capturing spatial patterns in the data while LSTM excel in capturing temporal dependencies in sequential data. In this approach the output of CNN is used to extract spatial features, while LSTM is used to capture temporal dependencies between the input data. The output of CNN and LSTM are then combined and fed into fully connected layer to make predictions or classify the data. Combining both the models allows to leverage the benefits of the models resulting in improved performance on task involving sequential data.

IV. IMPLEMENTATION

A. Implementation Environment

The study has been carried out on a computer equipped with a 11th Gen 3.20GHz Intel Core i5-11320H processor. All the experiments have been performed in Python. Integrated Development Environments (IDEs) like Jupyter Notebooks, Visual Studio Code were used. Cloud platforms like Google Collaboratory and Kaggle Kernel provided additional computing resources.Matplotlib and Seaborn were employed for data visualization, creating informative plots and graphs to analyze datasets. Flask will be used for user interface developme. Libraries and Frameworks like NumPy, Pandas, Scikit-learn are used for model building and training.

B. Result and Discussion

1) Evaluation Parameters used

a) Classification report tool: Classification report tool gives summary of the main classification metrics for each class of a model. It includes precision, recall, F1-score, and support for each class, as well as the weighted average of these metrics across all classes, macro average and accuracy.

- Precision: It measures the accuracy of positive predictions. It is given by the number of true positives divided by the sum of true positives and false positives.
- Recall: It is a measure of completeness of positive predictions. It is given by the ratio of number of true positives to the sum of true positives and false negatives.
- F1 score: It is the harmonic mean of precision and recall. It gives the balance between precision and recall.
- Support: It is the number of samples in each class.
- Accuracy: It measures how often the model correctly predicts the outcome. It is calculated by dividing the number of correct predictions to the total number of predictions.

b) Computational Efficiency(Training time): Training time is the time taken to train the model on a given dataset. It is a measure of how efficiently the model can learn from the data and optimize its parameters to make accurate predictions. Models that train quickly on larger datasets are more efficient.

c) Confusion Matrix: Confusion matrix shows how many correct and incorrect predictions are made by the model. It gives the count of true positives(TP), true negatives(TN), false positives(FP), false negatives(FN) for each class in the actual and predicted outcomes.

d) Learning curves :Learning curve represents the model's performance based on the size of the training data size. They help in understanding whether the model would benefit from additional data or if it has already converged.

e) AUC – ROC curve :Area Under the Receiver Operating Characteristics (AUC- ROC) curve is a graphical representation of performance of a classification model at different thresholds. ROC,Receiver Operating Characteristics plots true positive rate (TPR) vs false positive rates(FPR) at different thresholds. It is a representation of the effectiveness of the classification model. AUC, Area Under the Curve is the area under the ROC curve. It represents the capability of the model in classifying the different classes. The greater the AUC value better the models performance.

2) Analysis

The pre-processed data was split into 70% training data and 30% testing data. MLP, Extra-tree classifier and Catboost models are build for flood prediction. Fig 2.1 shows the learning curves for each of the model. Fig 2.2 represents the AUC-ROC curves for the models.

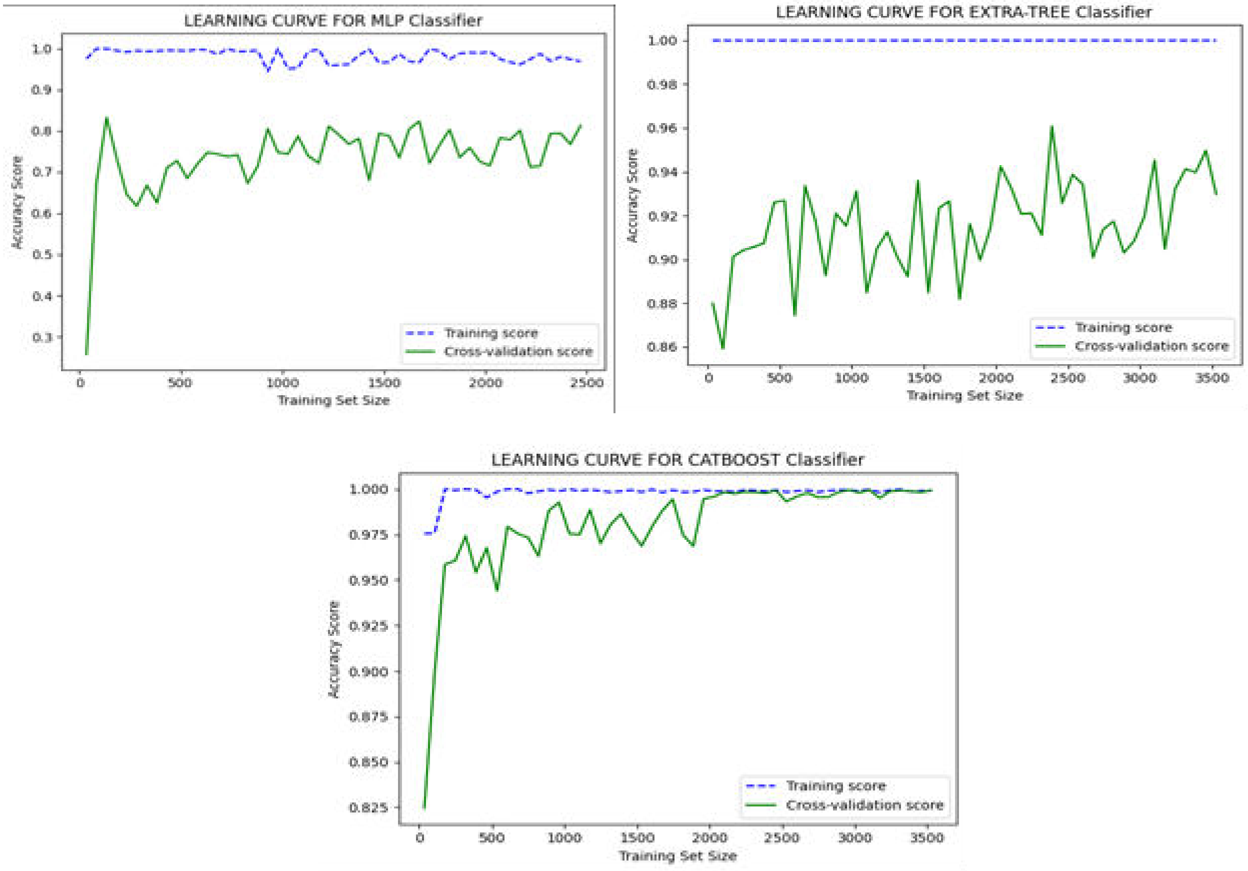
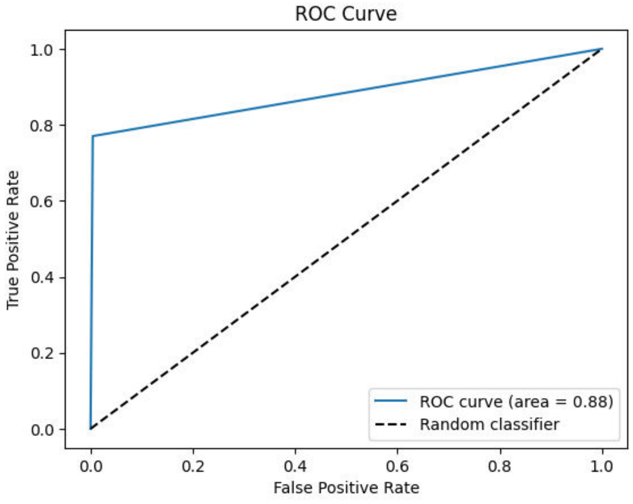
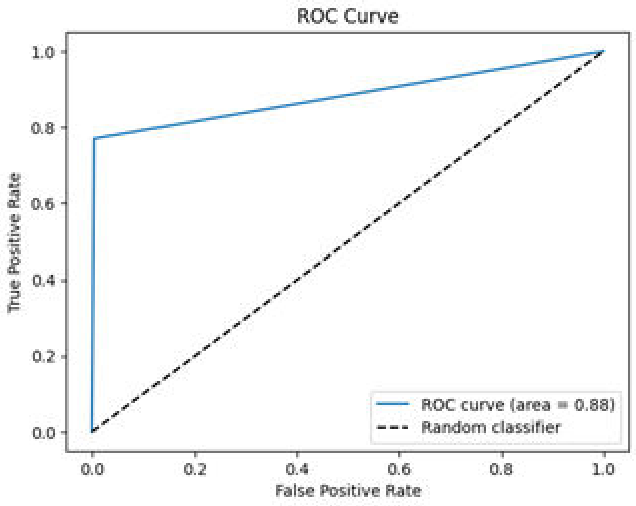


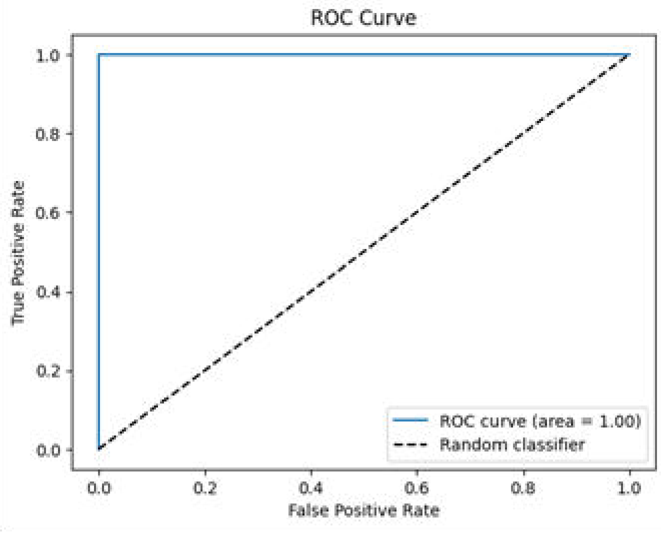
Figure 2. Learning curves



(a)



(b)



(c)

Figure 3. AUC-ROC curve for (a)MLP (b)Extra-tree (c)CatBoost

Table 2 Comparison of Classification report tool parameters

Model	Flood				No flood			
	Precision	Recall	F1 score	Support	Precision	Recall	F1 score	Support
MLP	0.657	0.932	0.771	74	0.996	0.969	0.982	1161
Extra tree	0.934	0.770	0.844	74	0.986	0.997	0.991	1161
CatBoost	1.000	1.000	1.000	74	1.000	1.000	1.000	1161

Table 3. Performance analysis

Parameters	MLP	Extra tree	Catboost
Accuracy	96.68	98.29	100
Precision	0.657	0.934	1
Recall	0.932	0.770	1
F1 score	0.770	0.844	1
Training time	0.179	0.0099	2.109

Table 4. Comparison of Confusion matrix parameter

Model	True Positive (TP)	True Negative (TN)	False Positive (FP)	False Negative (FN)
MLP	69	1125	36	5
Extra tree	57	1157	4	17
CatBoost	74	1161	0	0

Table 2 provides a comparison of Classification report tool parameters. Table 3 presents the performance analysis of the MLP, Extra tree and Catboost models. Table 4 provides a comparison of confusion matrix parameters.

V. CONCLUSION

In conclusion, our study delves into the realm of machine learning approaches for predicting floods and cyclones, addressing the need for accurate disaster forecasting. Through the utilization of various models such as MLP, Extra-Tree Classifier, Catboost, parallel CNN-LSTM , we aim to enhance predictive accuracy and minimize the detrimental impacts of these natural disasters on vulnerable regions. Our analysis incorporates a comprehensive evaluation of predictive performance using metrics including Learning Curves, ROC Curve and AUC Score, Classification Report, Computational Efficiency, and Confusion Matrix.The accuracy achieved by MLP is 96.68%, Extra tree is 98.29% and Catboost is 100%. Among the models used for flood prediction Catboost performed best with highest accuracy. In the future we aim to predict cyclone using parallel CNN-LSTM and integrate these models into a user-friendly interface to contribute to the advancement of disaster preparedness and decision-making processes for policymakers, emergency responders, and community leaders.

References

[1] K Sandhya Rani Kundra 1,B. Jaya Lakshmi 2,I V S Venugopal 3,Venkatesh Guthula 4, “Flood Prediction using MLP, CATBOOST and Extra-Tree Classifier”, International Journal of Recent Technology and Engineering (IJRTE), Volume-11,Issue-7s, 2023, https://doi.org/10.17762/ijritcc.v11i7s.6974.

[2] Marcel Motta 1, Miguel de Castro Neto 2, Pedro Sarmento 3,“ A mixed approach for urban flood prediction using Machine Learning and GIS”, International Journal of Disaster Risk Reduction-ElseVier, 2021, https://doi.org/10.1016/j.ijdrr.2021.102154.

[3] Vinothini A 1, Kruthiga L 2, Monisha U 3, “Predictionof Flash Flood using Rainfall by MLP Classifier”, International Journal of Recent Technology and Engineering (IJRTE), Volume-9, Issue-1,2020, 10.35940/ijrte.F9880.059120.

[4] P.Varalakshmi 1,N. Vasumathi 2,R. Venkatesan 3, “Tropical Cyclone prediction based on multi-model fusion across Indian coastal region”, Progress in Oceanography-Elsevier,2021, https://doi.org/10.1016/j.pocean.2021.102557.

[5] Abu Reza Md Towfiqul Islam 1, Swapan Talukdar 2, Susanta Mahato 3, “Flood susceptibility modelling using advanced ensemble machine learning models”, Geoscience Frontiers-Elsevier, 2020, https://doi.org/10.1016/j.gsf.2020.09.006.

[6] Haggag, M., Siam, A.S., El-Dakhakhni, W. et al. “A deep learning model for predicting climate-induced disasters”. Nat Hazards 107, 1009–1034 (2021), https://doi.org/10.1007/s11069-021-04620-0.

[7] Kim, M.; Park, M.-S.; Im, J.; Park, S.; Lee, M.-I. “Machine Learning Approaches for Detecting Tropical Cyclone Formation Using Satellite Data”. Remote Sens, 11, 1195,2020. https://doi.org/10.3390/rs11101195

[8] Y. Wu, X. Geng, Z. Liu and Z. Shi, "Tropical Cyclone Forecast Using Multitask Deep Learning Framework," in IEEE Geoscience and Remote Sensing Letters, vol. 19, pp. 1-5, 2022, Art no. 6503505, doi: 10.1109/LGRS.2021.3132395.

[9] Jishnu Saurav Mittapalli, Jainav Amit Mutha, Maheswari R, " An Intelligent Natural Disaster Predictor," in Energies Journal, vol. 16, pp.1459, 2020,https://doi. org/10.1007/s11069-019-03677-2.

[10] Chen R, Zhang W, Wang X., “Machine Learning in Tropical Cyclone Forecast Modeling: A Review.”, Atmosphere. 2020; 11(7):676. https://doi.org/10.3390/atmos11070676

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com