International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 7s
DOI: https://doi.org/10.17762/ijritcc.v11i7s.6974
***Article Received**: 30 March 2023 **Revised**: 12 May 2023 **Accepted**: 23 May 2023*

Flood Prediction using MLP, CATBOOST and Extra-Tree Classifier

[1] K Sandhya Rani Kundra, [2] B. Jaya Lakshmi, [3] I V S Venugopal, [4] Venkatesh Guthula
[1, 2, 3] Department of Information Technology, [4] Department of Computer Applications,
Gayatri Vidya Parishad College of Engineering (A), Visakhapatnam, Andhra Pradesh, India
e-mail: sandhyaranikk@gvpce.ac.in
e-mail: meet_jaya200@gvpce.ac.in
e-mail: venuillinda@gvpce.ac.in
e-mail: guthulavenkatesh699@gmail.com

Abstract— Flooding can be one of the many devastating natural catastrophes, resulting in the annihilation of life and damaging property. Additionally, it can harm farmland and kill growing crops and trees. Nowadays, rivers and lakes are being destroyed, and the natural water reservoirs are converted into development sites and buildings. Due to this, even just a bit of rain can cause a flood. To minimize the number of fatalities, property losses, and other flood-related issues, an early flood forecast is necessary. Therefore, machine learning methods can be used for the prediction of floods.

To forecast the frequency of floods brought on by rainfall, a forecasting system is built using rainfall data. The dataset is trained using various techniques like the MLP classifier, the CatBoost classifier, and the Extra-Tree classifier to predict the occurrence of floods. Finally, the three models' performances are compared and the best model for flood prediction is presented. The MLP, Extra-Tree, and CatBoost models achieved accuracy of 94.5%, 97.9%, and 98.34%, respectively, and it is observed that CatBoost performed well with high accuracy to predict the occurrence of floods.

Keywords- Flood Prediction, Machine Learning, CatBoost Classifier, Extra-Tree Classifier, MLP Classifier.

I. INTRODUCTION

A flood is a natural disaster that badly affects our lives. Generally, it is a regular phenomenon in India. It occurs primarily due to continuous heavy rainfall and the stagnation of water in an area for a long time. Global warming, deforestation, and increasing pollution are the indirect causes of floods. Several people lose their lives in floods, and lakhs of people are rendered homeless. Floods not only cause loss of life but also badly affect a country's economy.

Last year, many areas of India and other countries were prone to flooding. They are Nepal (12 September 2022), Uttar Pradesh (1 September 2022), Himachal Pradesh (21 August 2022), Andhra Pradesh (3 August 2022), Gujarat (12 July 2022), Manipur (1 July 2022), and Assam (21 June 2022).

There has been a lot of study done on flood prediction, but few approaches provide the estimated accuracy needed. Generally, machine learning techniques are more often used for the prediction because they offer fast and accurate results. In this work, MLP classifiers, CatBoost classifiers, and Extra- Tree methods are presented and analyzed to predict the floods.

A. Introduction to Problem Domain

A flood is an unnecessary overflow of water onto dry ground. Most of the floods are typically caused by heavy, continuous rainfall. The locations near the rivers are more likely to experience the flash floods. According to Jeerana Noymanee et al. [1], flooding is one of the most devastating problem. The authors have illustrated various misconceptions that are faced while facing the flood in real situations. To improve the flood prediction they have used hydrological modeling in conjunction with machine learning techniques.

Z. K. Lawal et al. [2] highlighted the benefits of using machine learning methods for getting alerts regarding floods, which results in reduction of loss caused by floods. The authors have used observational data and achieved high reliability with less computing power. The authors have demonstrated how decision tree worked better compared to the support vector classification in terms of accuracy.

As per, A. B. Ranit et al. [3], flooding is the state when a huge amount of water overflows onto a piece of land. In order to mitigate the risks caused by flooding due to climatic change, flood prediction using algorithms in machine learning is gaining insight about and enhances system scale. In their study, they used an Artificial Neural Network to estimate flood value in real time.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 7s
DOI: https://doi.org/10.17762/ijritcc.v11i7s.6974
***Article Received**: 30 March 2023 **Revised**: 12 May 2023 **Accepted**: 23 May 2023*

C. Kinage et al. [4], described floods as one of nature's most catastrophic tragedies, and are extremely difficult to model. On the gathered dataset, a variety of machine learning algorithms have been evaluated to see which algorithm performed the best and which parameters are most important. In their study, they also presented a machine learning based flood forecasting model and created an Android app for it. Miah Mohammad Asif Syeed et al. [5], supported the above findings, with the aid of several machine learning models, and their article intends to lower the risks associated with flooding while contributing to policy recommendations by making a precise prediction. To determine whether the model offered greater accuracy, a comparison on different metrics is done.

Thus, proposing a machine learning based method for prediction of floods is the need of the hour. Henceforth, in this work, various machine learning models are evaluated based on various performance metrics on the rainfall dataset and the best model is presented.

B. Objective

The main motivation behind this work is prediction of floods using machine learning techniques, and then calculate the performance of each technique based on the evaluation metrics and representing the best model for prediction. A rainfall dataset is considered and trained using machine learning techniques. The dataset contains some South Indian divisions and their different rainfall conditions that lead to floods.

The remaining paper is organized as follows: Section – II presents the current state of the art; Section – III analyses the dataset used; Section – IV presents the proposed framework; Section – V Emphasized on the obtained results; Section – VI concludes the work.

II. LITERATURE REVIEW

A variety of classification models like Decision Tree Induction, classifiers using Naïve Bayes (NB), logistic regression (LR), and Multi-layer perceptron (MLP) were proposed by Vinothini et al. [6]. The researchers have conducted comparative analysis of various classification systems in relation to various applications, and they examined several classification algorithms used for flood forecasting in their research. The main goal of their work is to give good information about the different classification methods used in flood forecasting and to make a better system for classifying floods.

Mohammed Khalf et al. [7] proposed a novel method using the ensemble model, to predict water level in relation to flood severity. They used the data collected from the sensor devices, and these values are passed as inputs to machine learning models to predict the severity of a flood.

J. Akshya et al. [8] conducted a lot of experiments using both unsupervised and supervised machine learning techniques and then used a combination of the two to predict floods. Their work developed a hybrid method to determine whether a region in an aerial photo has been flooded. Support Vector Machine (SVM) and k-means clustering have demonstrated high precision in identifying flooded areas, correctly classifying 92% of flooded images. The effectiveness of SVM is assessed by altering different kernel functions. According to the findings, a quadratic SVM can shorten the training and forecast times.

Floods can be predicted using a combination of deep learning and machine learning techniques like convolution neural networks and support vector machines as proposed by J.M.A Opella et al. [9]. Their study aims to create a precise flood risk and probability map using the data gathered from GIS (Geographical Information System) as well as current technological advancements. A feedforward neural network like ConvNet, which is good at processing images, is combined with SVM for prediction to get better results when mapping images. The output of the dilated convolution and deconvolution networks will be used as an input to create the final output of the SVM.

A.B. Ranit et al. [10] developed models for predicting floods in the future. The goal of forecast reliability is to give authorities and the general public early notice of an imminent flood. Flood forecasting (FF) is a challenging and difficult subject in hydrology. A flood forecasting method must give communities enough lead time to react. Forecasting skills in hydrology have risen, as have advances in knowledge for analysis and increases in data collection via satellite observations. This study examines different elements of flood forecasting, such as the models employed, methods for gathering inputs and displaying the results, and alerts.

Halit Enes Aydın et al. [11] developed flood susceptibility Maps using tree-based machine learning classifiers. Different machine learning models like LightGBM, CatBoost, XGBoost, and AdaBoost were evaluated using fourteen parameters and their research concluded that the models AdaBoost and LightGBM have highest accuracy. Their findings showed that flooding occurs mostly in places with lower heights, lower angles, proximity to banks of rivers, farming regions, and sparsely vegetated regions.

Thegeshwar Sivamoorthy et al. [12] proposed a Neural network approach to develop a flood forecast models that provided better performance and cost-effective solutions. To predict the occurrence of floods, the authors used MLP and a confusion matrix on a rain database. In order to capture different views on the given data they used various information metrics like Active recognition, deficit treatment, validation of data, along with information cleaning.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 7s
DOI: https://doi.org/10.17762/ijritcc.v11i7s.6974
***Article Received**: 30 March 2023 **Revised**: 12 May 2023 **Accepted**: 23 May 2023*

According to V.V. Ramalingam et al. [13], floods are erratic and challenging to forecast. The flood prediction structures have been enhanced by neural system designs, which have led to better execution and affordable solutions. Their study used rainfall datasets and neural network-based methods to estimate the likelihood of flooding. The accuracy calculation, confusion matrix identification, show how well their algorithms perform.

Finally, the work done by Parag Ghorpade et al. [14] is highly appreciable and inferred in this work who have done a review on forecasting the flood using machine learning methods. Their study discussed various notable algorithms used by experts to create solutions as machine learning algorithms have become more beneficial for flood predictions. They emphasised the advantages of computational models for flood modelling and the implications of data (such as water flow, rainfall, and humidity). Ainaa Hanis Zuhair et al. [15] worked in the similar direction and presented an overview of hybrid models in machine learning using datasets. They stated that hybridization, decomposition of the data, algorithmic ensemble, and optimization of the model as key strategies to improve the effectiveness of machine learning methods.

To summarize, many papers used different methods, algorithms, and techniques for flood prediction. The main objective of all the parallel researchers is to illustrate the most accurate model for flood prediction, which helps the public aware of the chances of a flood occurrence using machine learning techniques. Based on the outcomes of recent research attempts, the most appropriate machine learning techniques need to be used to achieve better results.

III. DATASET INTRODUCTION

The dataset contains the different rainfall conditions that led to floods in some urban areas in India from 1901 to 2015. Here different areas such as Kerala, the coastal regions of Karnataka and Andhra Pradesh, and the south and north interiors of Karnataka, Tamil Nadu, and Telangana are considered. The rainfall conditions are taken as the total amount of rainfall for each month from January to December. Season-wise rainfall of the first ten days of June is considered, which has the highest possibility of continuing rainfall that results in flooding in millimetres. The target column flood contains values 0's and 1's, in which the value "1" resembles floods that occurred in that year and value of "0" indicate that the flood did not occur that year.

The various features of the dataset used are shown in Fig.1.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
id	SUBDIVISI	YEAR	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC	ANNUAL	Jan-Feb	Mar-May	Jun-Sep	Oct-Dec	flood	10days_june
3887	KERALA	1901	28.7	44.7	51.6	160	174.7	824.6	743	357.5	197.7	266.9	350.8	48.4	3248.6	73.4	386.2	2122.8	666.1	0	274.8667
3888	KERALA	1902	6.7	2.6	57.3	83.9	134.5	390.9	1205	315.8	491.6	358.4	158.3	121.5	3326.6	9.3	275.7	2403.4	638.2	1	130.3
3889	KERALA	1903	3.2	18.6	3.1	83.6	249.7	558.6	1022.5	420.2	341.8	354.1	157	59	3271.2	21.7	336.3	2343	570.1	0	186.2
3890	KERALA	1904	23.7	3	32.2	71.5	235.7	1098.2	725.5	351.8	222.7	328.1	33.9	3.3	3129.7	26.7	339.4	2398.2	365.3	0	366.0667
3891	KERALA	1905	1.2	22.3	9.4	105.9	263.3	850.2	520.5	293.6	217.2	383.5	74.4	0.2	2741.6	23.4	378.5	1881.5	458.1	0	283.4
3892	KERALA	1906	26.7	7.4	9.9	59.4	160.8	414.9	954.2	442.8	131.2	251.7	163.1	86	2708	34.1	230	1943.1	500.8	0	138.3
3893	KERALA	1907	18.8	4.8	55.7	170.8	101.4	770.9	760.4	981.5	225	309.7	219.1	52.8	3671.1	23.7	328	2737.8	581.7	1	256.9667
3894	KERALA	1908	8	20.8	38.2	102.9	142.6	592.6	902.2	352.9	175.9	253.3	47.9	11	2648.3	28.8	283.7	2023.6	312.2	0	197.5333
3895	KERALA	1909	54.1	11.8	61.3	93.8	473.2	704.7	782.3	258	195.4	212.1	171.1	32.3	3050.2	65.9	628.3	1940.4	415.5	0	234.9
3896	KERALA	1910	2.7	25.7	23.3	124.5	148.8	680	484.1	473.8	248.6	356.6	280.4	0.1	2848.6	28.4	296.7	1886.5	637	0	226.6667
3897	KERALA	1911	3	4.3	18.2	51	180.6	990	705.3	178.6	60.2	302.3	145.7	87.6	2726.7	7.3	249.7	1934	535.7	0	330
3898	KERALA	1912	1.9	15	11.2	122.7	217.3	948.2	833.6	534.4	136.8	469.5	138.7	22	3451.3	16.9	351.1	2453.1	630.2	1	316.0667
3899	KERALA	1913	3.1	5.2	20.7	75.7	198.8	541.7	763.2	247.2	176.9	422.5	109.9	45.8	2610.8	8.3	295.2	1729	578.3	0	180.5667
3900	KERALA	1914	0.7	6.8	18.1	32.7	164.2	565.3	857.7	402.2	241	374.4	100.9	135.2	2899.1	7.6	215	2066.1	610.5	0	188.4333
3901	KERALA	1915	16.9	23.5	42.7	106	154.5	696.1	775.6	298.8	396.6	196.6	302.5	14.9	3024.5	40.4	303.1	2167	514	0	232.0333
3902	KERALA	1916	0	7.8	22	82.4	199	920.2	513.9	396.9	339.3	320.7	134.3	8.9	2945.3	7.8	303.4	2170.2	463.9	0	306.7333
3903	KERALA	1917	2.9	47.6	79.4	38.1	122.9	703.7	342.7	335.1	470.3	264.1	256.4	41.6	2704.8	50.5	240.4	1851.7	562.1	0	234.5667
3904	KERALA	1918	42.9	5	32.8	51.3	683	464.3	167.5	376	96.4	233.2	295.4	54.1	2501.9	47.9	767	1104.3	582.6	0	154.7667
3905	KERALA	1919	43	6.1	33.9	65.9	247	636.8	648	484.2	255.9	249.2	280.1	53	3003.3	49.2	346.8	2025	582.3	0	212.2667
3906	KERALA	1920	35.2	5.5	24.1	172	87.7	964.3	940.8	235	178	350.1	302.3	8.2	3303.1	40.6	283.7	2318.2	660.6	0	321.4333
3907	KERALA	1921	43	4.7	15	171.3	104.1	489.1	639.8	641.9	156.7	302.4	136.2	15.8	2719.9	47.8	290.3	1927.5	454.3	0	163.0333
3908	KERALA	1922	30.5	21.4	16.3	89.6	293.6	663.1	1025.1	320.6	222.4	266.3	293.7	25.1	3267.6	51.9	399.4	2231.2	585.1	0	221.0333
3909	KERALA	1923	24.7	0.7	78.9	43.5	80	722.5	1008.7	943	254.3	203.1	83.9	41.6	3484.7	25.3	202.3	2928.4	328.6	1	240.8333
3910	KERALA	1924	19.3	2.9	66.6	111	185.4	1011.7	1526.5	624	289.1	176.5	162.9	50.4	4226.4	22.2	363	3451.3	389.9	1	337.2333
3911	KERALA	1925	4.1	16.5	76.9	93.4	258.2	688.8	593.5	554.1	158.8	295.4	223.7	98.8	3062.1	20.5	428.5	1995.2	617.9	0	229.6
3912	KERALA	1926	28.6	5.8	23.1	55.8	222.6	563.9	885.2	536	322.7	216.7	88.8	16.2	2965.4	34.4	301.5	2307.8	321.7	0	187.9667
3913	KERALA	1927	18.8	35.3	49.6	86.5	265.4	720.2	888.2	315	335.6	135.8	137.6	6.8	2994.7	54.1	401.4	2258.9	280.2	0	240.0667
3914	KERALA	1928	12.7	65.9	51.3	121.1	81.9	590.7	420.6	553.2	75.9	321.5	155.2	52.7	2502.8	78.6	254.3	1640.4	529.4	0	196.9
3915	KERALA	1929	12.8	29.8	58.9	210.7	148	946.6	844	293.9	268.9	350.4	158.2	39.4	3361.6	42.6	417.6	2353.5	548	0	315.5333
3916	KERALA	1930	10.8	10.8	39	102.7	404.9	633.1	401.7	273.4	411.5	433.9	207	89.2	3018	21.6	546.5	1719.7	730.2	0	211.0333
3917	KERALA	1931	3.3	0.3	19.2	126.9	131.7	541.7	653.9	1199.2	163.2	149.3	164.3	106.5	3259.6	3.6	277.8	2558	420.1	1	180.5667
3918	KERALA	1932	0.1	19.3	28.6	113	646.5	341	716.4	423.2	317.3	543.2	223.2	31.3	3403	19.4	788.1	1797.8	797.7	0	113.6667
3919	KERALA	1933	1	9.3	36.9	139.5	738.8	859.3	773.4	479.5	469.7	397	126.1	42.3	4072.9	10.3	915.2	2581.9	565.5	1	286.4333
3920	KERALA	1934	74.5	1.7	47.7	92.4	106.7	852.9	415	337.2	48.4	335.9	93.4	4.9	2410.7	76.2	246.8	1653.5	434.2	0	284.3

Figure. 1 Dataset features

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 7s
DOI: https://doi.org/10.17762/ijritcc.v11i7s.6974
Article Received: 30 March 2023 Revised: 12 May 2023 Accepted: 23 May 2023

A. Dataset *Description*

The dataset contains a total of 804 entries.
The dataset contains 20 fields.
Training data: 562 entries (70%).
Testing data: 242 entries (30%).

B. *Data Preprocessing*

Data preprocessing is a method for transforming raw data into a structure that can be used and is effective. In this process, it includes data cleaning such as removing null values and label encoding.

- Data Cleaning: In order to make the prediction accurate, the null values are removed and are replaced by the mean value.
- Label Encoding: In order to convert text values into numeric values, label encoding is performed. Label encoding changes the text category into numeric integer values starting from 0. The values are assigned to each text item in alphabetical order. Here, by converting the subdivision column, the values for each area are as follows Kerala = 0, Andhra Pradesh Coastal Region = 1, Karnataka Coastal Region = 2, Karnataka North Interior Region = 3, Karnataka South Interior Region = 4, Tamil Nadu = 5 and Telangana = 6.

C. Data Splitting

Data splitting is the process that dissects the dataset into two entities. The model is trained with the first entity, and then tested with the second. The more the model is trained, the more accurate the results would be. Here, the dataset is divided in the ratio (70:30), meaning that 70 percent of the data, or 562 samples, are taken into consideration for training, and 30 percent of the data, or 242 samples, are taken into consideration for testing the model.

IV. PROPOSED FRAMEWORK

Fig. 2 illustrates the overall framework of the proposed system. A rain fall dataset is taken, and after the initial pre-processing, the data is dissected as training and testing sets. Then, data is trained on three models, namely MLP Classifier, Extra-Tree Classifier and CATBOOST classifier. Then, the models are evaluated on different evaluation performance metrics and the best model to predict the flood is presented.

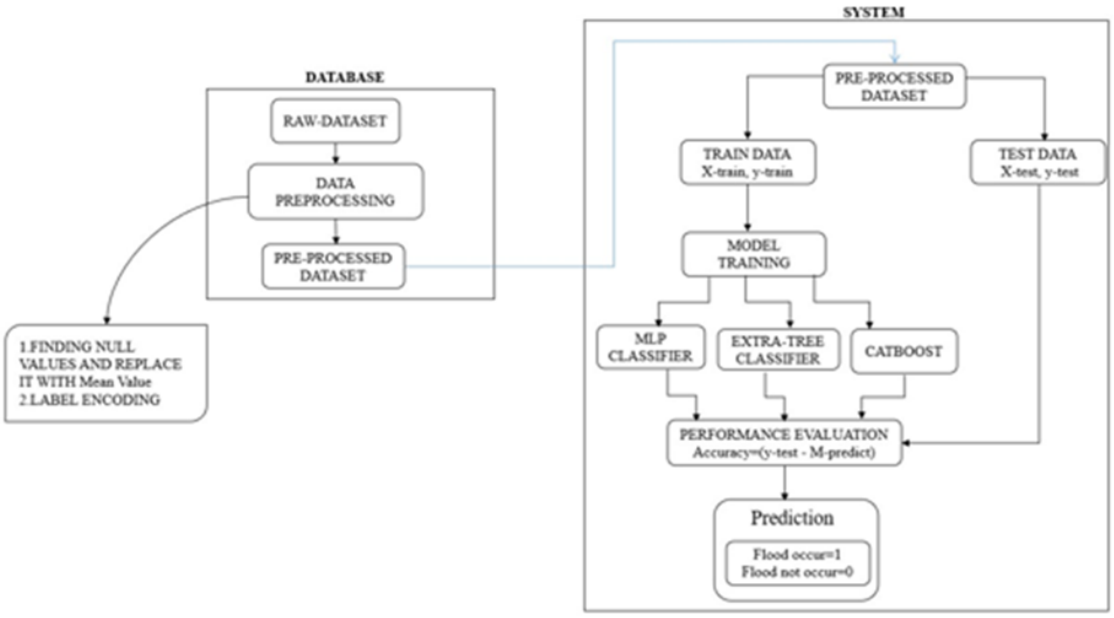

Figure. 2 System Architecture

The three models that are used in the framework are elaborated over here.

A. *MLP Classifier (Neural Networks)*

Acronym for MLP is Multi-layer Perceptron. As the name implies, it contains multiple layers as follows:

1. Input stratum
2. Hidden stratum
3. Output stratum

Example: Fig. 3 illustrates the process used by MLP that consists of multiple layers of interconnected nodes (also known as neurons) that can learn and process complex nonlinear relationships between input data and output predictions.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 7s
DOI: https://doi.org/10.17762/ijritcc.v11i7s.6974
Article Received: 30 March 2023 Revised: 12 May 2023 Accepted: 23 May 2023

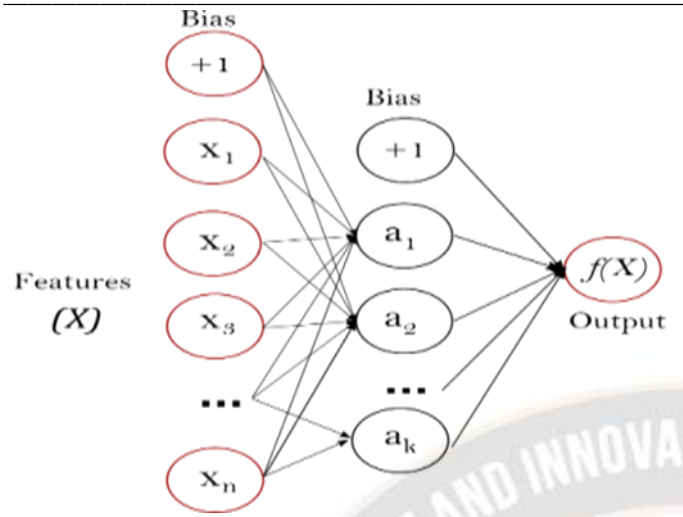

Figure. 3 MLP Process

Here x_1, x2….x_n are inputs, a_1, a2…. a_k are the hidden perceptron and f(x) is the output.

Working of MLP

MLP is a feed-forwarded neural network, so inputs are moved only in a forward direction. These inputs are multiplied by weights that are chosen at random, and each sum of multiplied values is linked to a hidden perceptron. Through an activated function, the perceptrons in the hidden layer can be made into a matrix as illustrated below.

Example:

a(x1*w11+x2*w13+b11)=h11
a(x1*w12+x2*w14+b22)=h12

Here, X-matrix represents input, and W-matrix represents random weights that generally lie in the range 0 to 1. b11, b22 are the bias, and H-matrix represents the generated perceptron's values in the hidden layer. Further, this process goes on up to the output layer and finally, the output is classified at the output layer.

B. CatBoost Classifier (Boosting)

The terms "category" and "boosting" are where the "CatBoost" moniker comes from. It performs exceptionally well with a wide variety of data types, including text, classification, numeric, and others. The first thing that happens when the data is provided, is that it randomizes the order of the data and then performs target encoding for the categorical column relative to the target column.

CatBoost is a gradient boosting library that can handle categorical features directly, without the need for one-hot encoding. One of the techniques it uses to handle categorical features is target encoding. Since this library was modeled after a gradient-boosting library, the word "Boost" derives from the term "gradient- boosting". It is an iterative approach. Iterations are performed until the loss is reduced.

Target encoding is a machine learning approach that transforms categorical features into numerical features, by replacing each categorical value with the mean (or another aggregation function) of the target variable for that category. CatBoost is a popular gradient boosting library that provides an implementation of target encoding.

Example: Fig. 4 shows how target encoding is performed using CatBoost.

Month	City	Flood
5	HYD	1
8	BNG	1
10	BNG	1
11	VZG	0
12	BNG	1

Month	City	E_value	Flood
5	HYD	0.5	1
8	BNG	0.5	1
10	VZG	0.5	1
11	VZG	0.25	0
12	BNG	0.75	1

Figure. 4 CatBoost Target Encoding

Encoded_value =(current_count +prior)/Max_Count +1

current_count = The sum of the target value for that category

$$\begin{bmatrix} X1 \\ X2 \end{bmatrix} \begin{bmatrix} W11 & w12 \\ W13 & w14 \end{bmatrix} = \begin{bmatrix} h11 \\ h12 \end{bmatrix}$$

feature (up to the current one)

prior = It's a constant value. It is determined by avg (target of category types).

Max_Count = Total number of categorical observations (up to the current one).

The steps used by CatBoost classifier are elaborated here:

Load the dataset – [Xi, Yi]

Step-1: Train the model m1 with data, calculating the error1= [Yi- Y_pred] .f1 (xi).

Step-2: Now train model m2 with [Xi, error1], calculating the error2= [Yi-Y_pred]. f2 (xi).

Step-3: At the end of step 2, calculate F1 (Xi) =f1(xi)+f2(xi)

Step-4: Now train model F1 [Xi, error2], calculating the error3= [Yi- F1(xi)].

Step-5: Now train model m3 with [Xi, error3], f3 (xi) =error.

At the end of step 5, F2=f1 (xi) +f2 (xi) +f3 (xi)

The steps will be iterated until the loss is reduced.

Final model, $F_m(x) = F_{m-1}+$

Fig. 5 shows the various iterations of the CatBoost process.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 7s
DOI: https://doi.org/10.17762/ijritcc.v11i7s.6974
***Article Received**: 30 March 2023 **Revised**: 12 May 2023 **Accepted**: 23 May 2023*

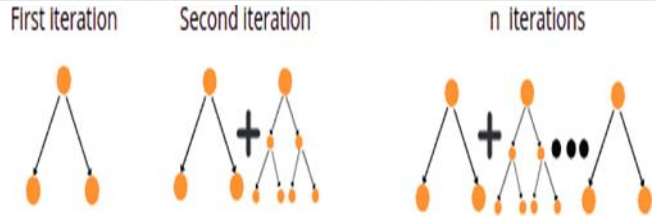

Figure. 5 CatBoost Process

C. Extra-Tree Classifier (Bagging)

An example of an ensemble method is the Extra-Tree classifier, which outputs its classification based on the aggregated results of numerous decision trees that have been gathered together in a "forest". In the Extra-tree classifier, random splitting of decision trees is done. So, it is called "Extremely Randomized Trees". For the working of the Extra-Tree algorithm, from the training dataset, a huge volume of unpruned decision trees are created. Predictions are done in classification by majority voting from decision trees.

Steps in Extra-Tree Classifier:

Step1: Random selection of an input.

Step2: Use random vectors to build multiple decision trees.

Step3: Combine the decision trees.

Step4: Prediction of the result.

Fig. 6 shows the Extra-Tree Classifier method used for classification problems that builds a collection of decision trees and combines their predictions to make final predictions.

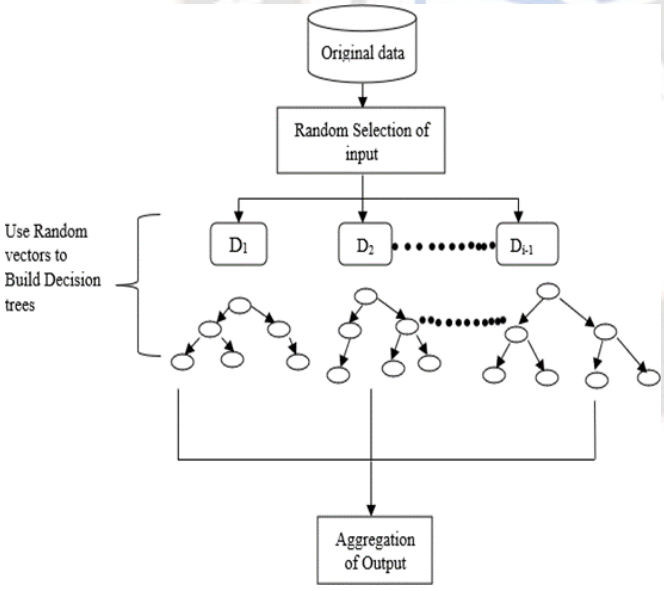

Figure. 6 Extra-Tree Classifier Process

V. RESULTS AND ANALYSIS

The performance of the classification models presented in the proposed framework, are evaluated and compared using the Learning curves, ROC and AUC graphs, and classification report tools, which are elaborated in the subsequent sections.

A. Learning Curves

The Learning curve illustrates the changes in the error metric value as the size of the training set increases during the training and validation phases. Learning curves are a way to visualize the efficiency of machine learning models as amount of training data increases. They plot the training and validation accuracy (or loss) based on the quantity of training instances. Learning curves helps in diagnosing whether a model can be an overfit or underfit to the data, and can provide guidance on whether collecting more training data would be beneficial.

In this work, 70% of the data is used as the training set, and 7 folds of cross validation are used for testing.

- Training data: 70% = 562 entries.
- Cross-validation CV = 7 folds.

a) MLP Classifier Learning Curve

```
MLP_specificity: 0.985

<Figure size 432x288 with 0 Axes>
```

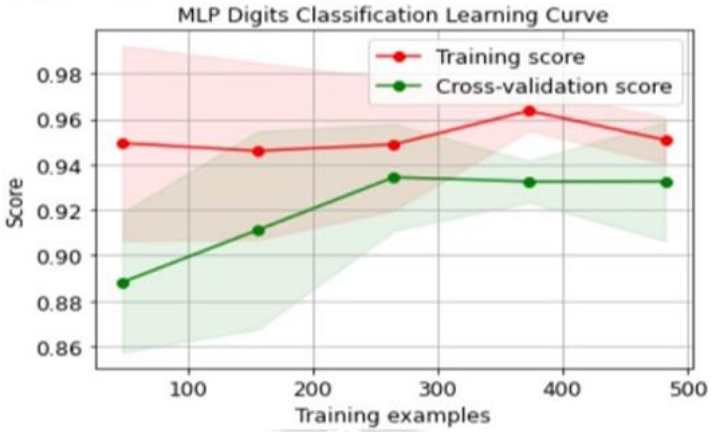

Figure. 7 MLP Learning Curve

From the above Fig.7, x-axis and y-axis pertain to the quantity of training examples, and accuracy score respectively. The red line is an indicator for the training score of the MLP classifier at different sizes of training examples (0 to 562 entities). The green line indicates a cross-validation score of 7. From Fig.7 it can be inferred that, the error value for the MLP classifier is nearly 0.02%, and when the training examples are increased, the MLP may underfit the dataset.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 7s
DOI: https://doi.org/10.17762/ijritcc.v11i7s.6974
***Article Received**: 30 March 2023 **Revised**: 12 May 2023 **Accepted**: 23 May 2023*

b) Extra-Tree Classifier Learning Curve

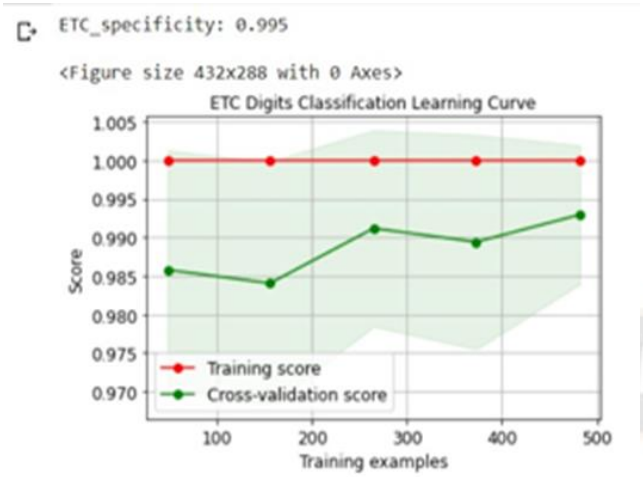

Figure. 8 Extra-Tree Learning curve

From the above Fig.8, x-axis and y-axis pertain to the quantity of training examples, and accuracy score respectively. The red line is an indicator for the training score of Extra- Tree classifier at different sizes of training examples from (0 to 562 examples). The green line indicates a cross-validation score of 7. From Fig.8 it can be inferred that, the error value for the Extra- Tree classifier is nearly 0.006%.

c) CatBoost Classifier Learning Curve

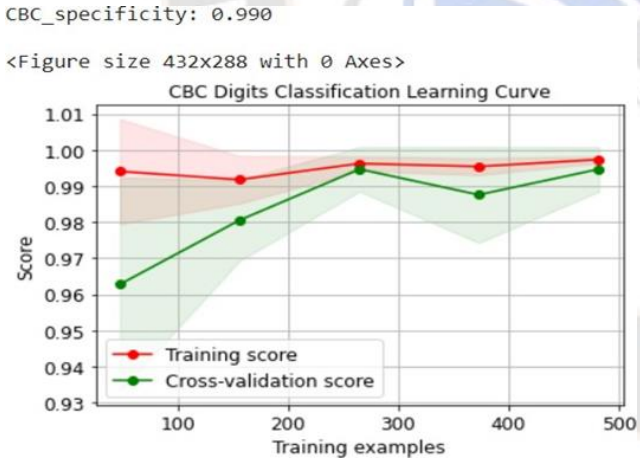

Figure. 9 CatBoost Learning Curve

From the above Fig.9, x-axis and y-axis pertain to the quantity of training examples, and accuracy score respectively. The red line is an indicator for the training score of the CatBoost classifier at different sizes of training examples, from (0 to 562 examples). The green line indicates a cross- validation score of 7. From Fig.9 it can be inferred that, the error value for the CatBoost classifier is nearly zero. When the number of training examples are increased, CatBoost minimized the error value.

d) *Comparative Analysis of Learning Curves of the three models*

By comparing the three graphs Fig.7, 8 & 9 respectively, the conclusion can be drawn that CatBoost is the most generalized model for the flood data when compared with MLP and Extra-Tree classifier.

B. ROC and AUC Graphs

Utilizing Receiver Operating Characteristic (ROC) and Area under the Curve (AUC) techniques, the binary classification models are assessed. ROC curves plots the true-positive rate (TPR) versus the false-positive rate (FPR) at various thresholds of classification. The AUC represents the degree or measure of separability between the positive and negative classes, and it measures the overall performance of a binary classification model over different thresholds of classification. The AUC can range in between from 0 to 1, when a models AUC score is 0.5 performs just as poorly as guessing at random and a model with an AUC score of 1.0 perfectly separates the positive and negative classes. Therefore, the model's capacity to differentiate between positive and negative samples is improved by a higher AUC value.

a) MLP ROC and AUC Graph

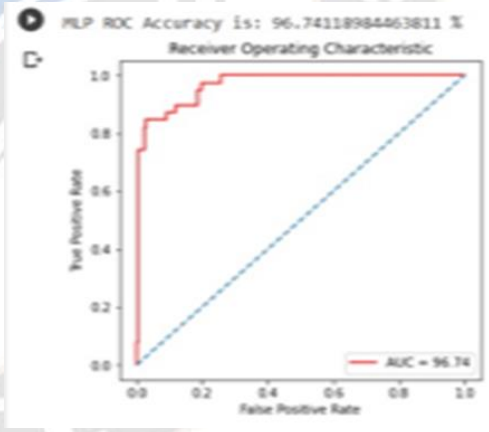

Figure. 10 MLP ROC Graph

In the above Fig.10, the y-axis and x-axis pertain to the TP rate and FP rate scores, respectively, at various decision rules (thresholds). The red line (ROC curve) indicates the attainment of the MLP classifier at different decision rules. The blue line indicates the linear classification. The ROC score is calculated by measuring area beneath ROC curve. Here, MLP has a ROC value of is 96.74%.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 7s
DOI: https://doi.org/10.17762/ijritcc.v11i7s.6974
***Article Received**: 30 March 2023 **Revised**: 12 May 2023 **Accepted**: 23 May 2023*

b) EXTRA-TREE ROC and AUC Graph

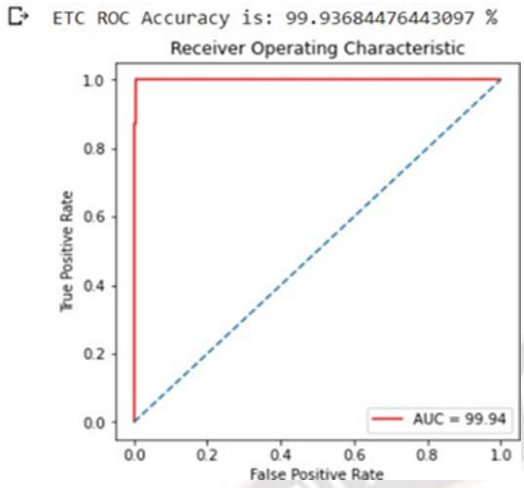

Figure. 11 Extra-Tree ROC Graph

In the above Fig.11, the y-axis and x-axis pertain to the TP rate and FP rate scores, respectively, at various decision rules (thresholds).The red line (ROC curve) indicates the performance of the Extra-Tree classifier model at different decision rules. The blue line indicates the linear classification. The ROC score is calculated by measuring distance area beneath the ROC curve. Here, the ROC value of Extra-Tree is 99.9%.

c) *CatBoost ROC and AUC Graph*

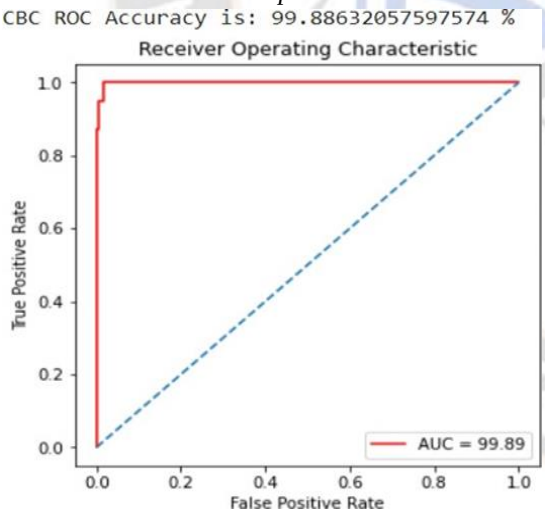

Figure. 12 CatBoost ROC Graph

In the above Fig.12, the y-axis and x-axis pertain to the TP rate and FP rate scores, respectively, at various decision rules (thresholds).The red line (ROC Curve) indicates the performance of the model, CatBoost classifier at different decision rules. The blue line indicates the linear classification. The ROC score is calculated by measuring area beneath ROC curve. Here, CatBoost has a ROC value of 99.8%.

d) *Comparative Analysis of ROC and AUC graphs of the three models*

By comparing the three graphs Fig. 10, 11, and 12 respectively, it can be inferred that the Extra-Tree classifier produced fewer false positive results when compared with the CatBoost and MLP classifier.

C. Classification Reports

Classification reports provide a compendious of the performance of a machine learning classification model on a per-class basis. They typically include metrics such as F1 score, recall, precision, and support. The ratio of true positives versus the total number of predicted positives is called as the Precision. It measures the accuracy of accurate predictions. The proportion of real positives to the total number of real positives is known as recall. It is a measure of the completeness of accurate predictions, and F1 score is the harmonic mean of recall and precision. Support is the total number of samples in every class. Classification reports can help diagnose whether a model is biased towards certain classes, and can provide guidance on which classes to focus on for improving model performance.

Accuracy: It evaluates the proportion of the accurate predictions among all samples.

Accuracy = TN+TP/(FP+TP+FN+TN)

Precision: It represents the percentage of accurate positive forecasts among all of the model's positive predictions.

Precision = TP/(FP+TP)

Recall: It measures the percentage of true positive forecasts among all actual positive samples.

Recall = TP/(FN+TP)

F1 Score: It is the harmonic mean of recall and precision, balancing the trade-off between the recall and precision.

F1-score =2(recall*precision)/(recall+precision)

Table 1. Performance of the model

Model	*Accuracy*	*ROC*	*Recall*	*Precision*	*F1-Score*
MLP	94.6	96	93	96	95
EXTRA-TREE	97.9	99	98	95	98
CATBOOST	98.3	99	97	97	98

From the Table 1, it could be inferred that the CatBoost produced maximized performance at all performance metrics namely Precision, ROC, Recall and accuracy when compared with the Extra-Tree Classifier and the MLP Classifier.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 7s
DOI: https://doi.org/10.17762/ijritcc.v11i7s.6974
***Article Received**: 30 March 2023 **Revised**: 12 May 2023 **Accepted**: 23 May 2023*

a) *Accuracy Comparison of the three models*

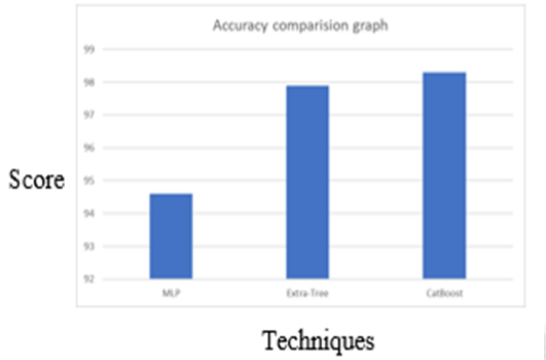

Figure. 13 Accuracy Comparison

The above graph Fig. 13 shows a comparative analysis of three models: MLP, Extra-Tree, and CatBoost Classifier. Here, the x-axis is an indicator of the techniques or models used, and the y-axis an indicator of the accuracy score. The accuracies for MLP, Extra- Tree, and CatBoost Classifier are nearly 94.5, 97, and 98 percent respectively. So CatBoost produces more accurate results than the other two models.

VI. CONCLUSION

In this study, the various models for flood prediction are analysed and the best model is presented. Analysis is done using rainfall statistics from some areas of India. The dataset has been trained with the MLP classifier, the Extra-Tree classifier, and the CatBoost classifier. The models MLP, Extra-Tree, and CatBoost achieved accuracy of 94.5%, 97.9%, and 98.34%, respectively. Henceforth, it can be concluded that among the three models, CatBoost performed well with high accuracy to predict the occurrence of flood.

VII. FUTURE WORK

Other applications of Artificial Intelligence such as deep learning can be utilized for obtaining more accuracy. More conditions and features for flood occurrence can be analyzed and a model can be developed.

REFERENCES

[1] Jeerana Noymanee, Thanaruk Theeramunkong, “Flood Forecasting with Machine Learning Technique on Hydrological Modeling”, Procedia Computer Science, Volume 156, 2019, pp. 377-386.

[2] Ghazaly, N. M. . (2020). Secure Internet of Things Environment Based Blockchain Analysis. Research Journal of Computer Systems and Engineering, 1(2), 26:30. Retrieved from https://technicaljournals.org/RJCSE/index.php/journal/article/view/8

[3] Z. K. Lawal, H. Yassin and R. Y. Zakari, "Flood Prediction Using Machine Learning Models: A Case Study of Kebbi State Nigeria," 2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), Brisbane, Australia, 2021,pp.1-6,doi: 10.1109/CSDE53843.2021.9718497.

[4] A. B. Ranit and P. V. Durge, "Flood Forecasting by Using Machine Learning," 2019 International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 2019,pp.166-169,doi: 10.1109/ICCES45898.2019.9002579.

[5] C. Kinage, A. Kalgutkar, A. Parab, S. Mandora and S. Sahu, "Performance Evaluation of Different Machine Learning Based Algorithms for Flood Prediction and Model for Real Time Flood Prediction," 2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA), Pune, India, 2019, pp. 1-7, doi: 10.1109/ICCUBEA47591.2019.9128379.

[6] Dr. Bhushan Bandre. (2013). Design and Analysis of Low Power Energy Efficient Braun Multiplier. International Journal of New Practices in Management and Engineering, 2(01), 08 - 16. Retrieved from http://ijnpme.org/index.php/IJNPME/article/view/12

[7] Asif Syeed, Miah & Farzana, Maisha & Namir, Ishadie & Ishrar, Ipshita & Nushra, Meherin & Rahman, Tanvir, “Flood Prediction using Machine Learning models ”, 2022, doi:10.48550/arXiv.2208.01234.

[8] Vinothini A, Kruthiga L, Monisha U, “Prediction of Flash Flood Using Rainfall by MLP Classifier,” International Journal Of recent Technology and Engineering (IJRTE), ISSN:2277-3878,Volume-9 Issue-1, May 2020.

[9] Mohammed Khalf, Haya Alaskar, Abir Jaafar Hussian et al, “IOT- Enabled Flood Severity Via Ensemble Machine Learning Models,” Institute of Electrical and Electronics and Engineering(IEEE), Digital Object Identifier 10.1109/ACCESS.2020.2986090.

[10] J. Akshaya and P. L. K. Priyadarshini, "A Hybrid Machine Learning Approach for Classifying Aerial Images of Flood-Hit Areas," 2019 International Conference on Computational Intelligence in Data Science(ICCIDS), 2019,pp.1-5,Doi: 10.1109/ICCIDS.2019.8862138.

[11] J. M. A. Opella and A. A. Hernandez, "Developing a Flood Risk Assessment Using Support Vector Machine and Convolutional Neural Network: A Conceptual Framework," 2019 IEEE 15th International Colloquium on Signal Processing & Its Applications (CSPA), 2019, pp. 260-265, Doi: 10.1109/CSPA.2019.8695980.

[12] A. B. Ranit and P. V. Durge, "Different Techniques of Flood Forecasting and Their Applications," 2018 International Conference on Research in Intelligent and Computing in Engineering (RICE), 2018, pp. 1-3, Doi: 10.1109/RICE.2018.8509058.

[13] Pérez, C., Pérez, L., González, A., Gonzalez, L., & Ólafur, S. Personalized Learning Paths in Engineering Education: A Machine Learning Perspective. Kuwait Journal of Machine Learning, 1(1). Retrieved from http://kuwaitjournals.com/index.php/kjml/article/view/107

[14] Halit Enes Aydin & Muzaffer Can Iban, 2023. "Predicting and analyzing flood susceptibility using boosting-based ensemble machine learning algorithms with SHapley Additive exPlanations," Natural Hazards: Journal of the International

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 7s
DOI: https://doi.org/10.17762/ijritcc.v11i7s.6974
***Article Received**: 30 March 2023 **Revised**: 12 May 2023 **Accepted**: 23 May 2023*

Society for the Prevention and Mitigation of Natural Hazards, Springer;International Society for the Prevention and Mitigation of Natural Hazards, vol. 116(3), pages 2957-2991, December-2022.

[15] Thegeshwar Sivamoorthy, Asif Mohammed Ansari, Dr. B. Sivakumar, V. Nallarasan, "Flood Prediction Using ML Classification Methods on Rainfall Data", International Journal For Research in Applied Science and Engineering Technology, Volume 10 Issue 4 Apr 2022.

[16] Maria Gonzalez, Machine Learning for Anomaly Detection in Network Security , Machine Learning Applications Conference Proceedings, Vol 1 2021.

[17] Dr. V.V. Ramalingam, Rohan Mishra, Sagar Parashari, "Prediction of Flood by Rainfall using MLP Classifier of Neural Network Model", International Journal of Advanced Science and Technology, 2020, Volume 29 Issue 06, pp.2804 - 2812

[18] P. Ghorpade et al., "Flood Forecasting Using Machine Learning: A Review," 2021 8th International Conference on Smart Computing and Communications (ICSCC), Kochi, Kerala, India, 2021, pp. 32-36 , doi: 10.1109/ICSCC51209.2021.9528099.

[19] Ainaa Hanis Zuhairi,Fitri Yakub,Sheikh Ahmad Zaki,Mohamed Sukri Mat Ali,"Review of flood prediction hybrid machine learning models using datasets",The 9th AUN/SEED-Net Regional Conference on Natural Disaster, 2022, doi:10.1088/1755-1315/1091/1/012040

[20] Danso-Amoako et al. "Predicting dam failure risk for sustainable flood retention basins:" A generic case study for the wider greater manchester area. Comput. Environ. Urban Syst. 2012, 36, 423–433.

[21] Kumar , D. A. ., & Das, S. K. . (2023). Machine Learning Approach for Malware Detection and Classification Using Malware Analysis Framework. International Journal of Intelligent Systems and Applications in Engineering, 11(1), 330–338. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/2543

[22] Thakre, B., Thakre, R., Timande, S., & Sarangpure, V. (2021). An Efficient Data Mining Based Automated Learning Model to Predict Heart Diseases. Machine Learning Applications in Engineering Education and Management, 1(2), 27–33. Retrieved from http://yashikajournals.com/index.php/mlaeem/article/view/17

[23] F. R. G. Cruz, M. G. Binag, M. R. G. Ga and F. A. A. Uy, "Flood Prediction Using Multi-Layer Artificial Neural Network in Monitoring System with Rain Gauge, Water Level, Soil Moisture Sensors," TENCON 2018 - 2018 IEEE Region 10 Conference, 2018, pp. 2499-2503, DOI: 10.1109/TENCON.2018.8650387.

[24] G. Kaur and A. Bala, "An Efficient Automated Hybrid Algorithm to Predict Floods in Cloud Environment," 2019 IEEE Canadian Conference of Electrical and Computer Engineering (CCECE), 2019, pp. 1-4, Doi: 10.1109/CCECE.2019.8861897.

[25] Amir Mosavi , Pinar Ozturk and Kwok-wing Chau , "Flood Prediction Using Machine Learning Models," Literature Review by Department of Computer Science (IDI), Norwegian University of Science and Technology (NTNU), Trondheim, NO-7491, Norway.

[26] F. A. Ruslan, K. Haron, A. M. Samad and R. Adnan, "Multiple Input Single Output (MISO) ARX and ARMAX model of flood prediction system: Case study Pahang," 2017 IEEE 13th International Colloquium on Signal Processing & its Applications (CSPA), 2017, pp.179-184,Doi:10.1109/CSPA.2017.8064947