How to Cite:
Natarajan, K., Gangaraju, R., & Tumati, A. B. (2022). Hybrid ML and DL models for flood level prediction. *International Journal of Health Sciences*, *6*(S9), 3699–3708. Retrieved from https://sciencescholar.us/journal/index.php/ijhs/article/view/13434

Hybrid ML and DL models for flood level prediction

Karthika Natarajan
Assistant Professor, School of Computer Science and Engineering, Vellore Institute of Technology, AP.
Email: karthika.n@vitap.ac.in

Rohith Gangaraju
Graduate Student, Presently PursuingBachelor Of Technology in computer science engineering atVellore Institute of Technology, AP.
Email: rohith2742@gmail.com

Akhilesh Babu Tumati
Graduate Student, Presently Pursuing Bachelor Of Technology in computer science engineering at Vellore Institute of Technology, AP.
Email: akhileshbabutumati2002@gmail.com

***Abstract*---**Flood occurrences happen in a particular place(that particular place may be anywhere) due to many causes, it may be due to dam's outbreak, because of poor infrastructure design aid during heavy rainfall times, Massive Rainfalls in a particular area, snowmeltin northern areas, and climate changes. In this work, Heavy Rainfall is considered as the major factor or key attribute for predicting the floods, particularly, the rainfall downpour amount for each month and annual rainfall value to predict flood occurrence in a particular place and in that particular season. In the work, Artificial models namely KNN Classifier, Logistic Regression, Decision Tree Classifier, Random Forest Classifier and Ensemble Learning and Deep learning architectures i.e. LSTMs (Long-Short Term Memory) and MLP (Multi Layer Perceptron) for predicting floods by comparing the accuracy obtained by ML models and Deep Learning models. The analysis is going to be done in three parts, the first part is the comparison of testing and training accuracies between each and every Machine learning models i.e. 5 state-of-the-art models which are mentioned above, the second part is to compare the test and train accuracies between the two deep learning architecture which are mentioned above. Finally, considering all the 7 architectures, identify the best model which is suitable for predicting the floods in a particular area.

Manuscript submitted: 9 May 2022, Manuscript revised: 18 July 2022, Accepted for publication: 27 August 2022

***Keywords*---**Flood Prediction, Artificial Models, Machine Learning Models, Deep Learning Models, Accuracy.

Introduction

Floods are hazardous in many ways, floods can occur in a form of natural disaster(heavy rainfall, river overflow, etc) or in a form of infrastructural error like break out of dams or flood gates, etc., we all know that floods create so much loss for people residing in the coastal areas or in other terrains [13]. Due to flood occurrence, mainly the farmers as well as common people, the crops which they have been harvesting for some period of time will be damaged in an instant. There is a susceptibility of flooding in that particular place or in other terrains it might cause landslides which later becomes a serious hazard [12]. So many state-of-the-art Machine Learning Models which were teamed up with sensors i.e. water-level readings will predict the susceptibility of floods in a particular area. The aim of this paper is to apply different state-of-the-art models to predict the flood (it might be any place) and analyze the state-of-the-art-models[10].

Related Work

This paper has examined different Machine Learning [5] and deep learning models [15]. They are mentioned below.

a) KNN Classifier: It is a distance based machine learning algorithm. It has the power for classification and regression of data points as well since it is a supervised learning algorithm. If we take a new flood susceptible point from a dataset, it first calculates the distances from its neighbors, then it picks the nearest one out of them. It is advisable to pick an odd number of datapoints of different classes in order to classify them with ease.This works on the rule of majority votes in the near neighbors.

b) Logistic Regression: It is another supervised machine learning algorithm that can be used for classification as well as for class probability elimination [3].Logistic regression is a linear regression used for classification problems.Logistic regression is used to classify different items. If the classification is to be made between two classes then it is a binary logistic regression. In this type of regression there are possibly 2 types of outcomes either 0 or 1. When there are more classes to be classified, it is said to be multinomial logistic regression.

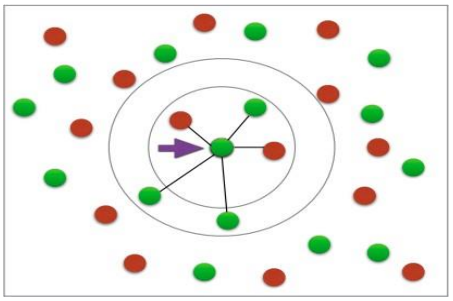

Figure 1 - knn algorithm classification

In this type of regression, they can have any number of outcomes, i.e.,the model can give any class as the output. When all the classes are ordered, then it is said to be ordinal logistic regression.Output of this type of regression is same as multinomial logistic regression since it is used for binomial classification when the dataset is categorical. In logistic regression, we do linear regression by applying a sigmoid function for classification.

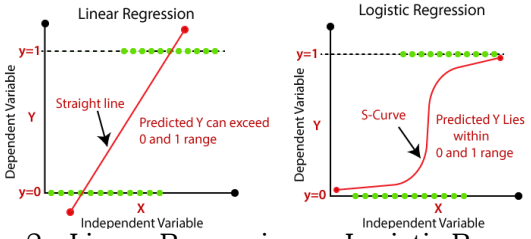

Figure 2 - Linear Regression vs Logistic Regression

c) Decision Tree Classifier: Decision Tree Classifier is a supervised learning based algorithm which is used for both classification and regression problems. This model's aim is to predict the class or a value of the target variable by following the decision tree algorithm from the training data. In the structure, internal nodes of a tree represent features of data, leaf nodes represent outcome and decisions are represented by branches.This decision tree uses multiple binary classification to get the result out as is represented in Figure 3.

d) Random Forest: Random Forest is a supervised learning based algorithm which is widely used in both regression and classification problems [11]. This model builds decision trees on different samples. The majority vote is taken for classification and the average value is taken for regression problems respectively.

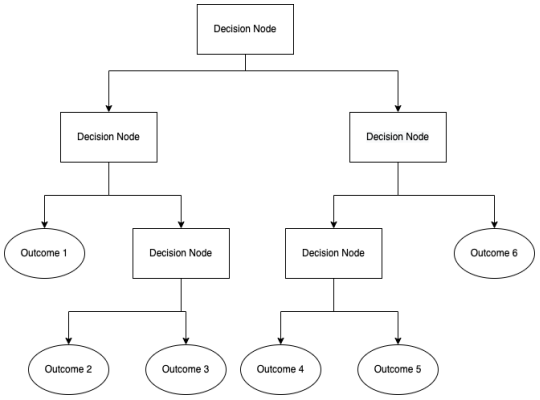

Figure 3- Model of decision tree classifier

Speed of the model is very high due to the architecture of classification. In a random forest, 'n' number of records are made and each record will be assessed individually by building a decision tree. Each decision tree will give a single output and the final output will be decided by the maximum number of votes from these outputs.A sample dataset is divided into 3 decision trees. Each decision tree predicts whether the test data belongs to class A or class B. If the majority of trees predict the test data belongs to class A then the random forest gives output as test data belongs to class
A. The same is shown in Figure 4.

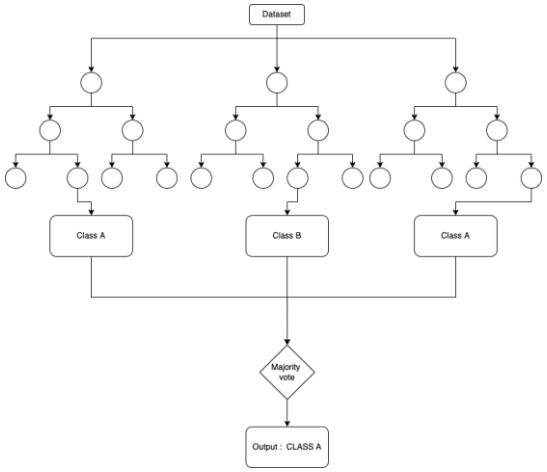

Figure 4 - Random Forest Working architecture

e) Ensemble Learning: Ensemble learning usually uses multiple models to predict the class of the test data. The random forest predicts the output from different decision trees. Here multiple models come into the picture instead of decision trees. Models can be of the same type or different. All models in ensemble learning are called Ensemble members. All these members must be trained on the same data. Ensemble learning is a supervised-based learning algorithm. This model combines KNN, Logistic Regression, Decision tree,

Random Forest predictive analysis to make another predictive analysis to obtain better accuracy Which is illustrated in Figure 5.

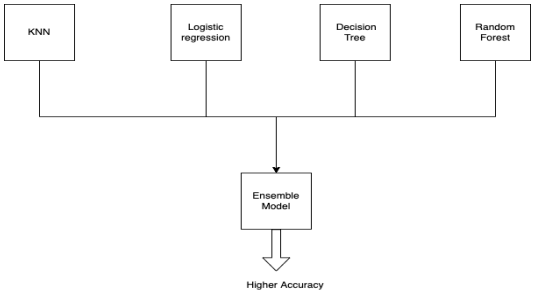

Figure 5 - Ensemble model of Machine Learning models in this paper

Now coming to the DeepLearning models [4], in this paper, two state-of-the-art deep learning models are considered. They are LSTM(Long Short Term memory) neural networks and MLP(Multi Layered Perceptron) neural networks [14].

f) LSTM: Recurrent Neural Network (RNN) has short-term memory due to vanishing gradient problems which resulted in the introduction of LSTM [1]. This is achieved by introducing other state Long-term memory. Long-short term memory(LSTM) is a type of recurrent neural network. LSTM has 2 states a) Hidden state - Short term memory b) Cell state - Long term memory. Long term and short term memory together remember the data for further use of the data and are also capable of learning order dependence. Due to the ability to read or memorize the long-term dependencies, LSTM works on any dataset if data is sequential. The prediction is done by taking the previous information/data into consideration [7][8].

e) MLP : Multi Layer Perceptron can remember a large amount of sequential data [6]. RNN's are used for remembering data as the patterns are good at remembering the historical sequential or related data[2]. This depends on the neural network for classification of the testing data. Same is represented in Figure 6

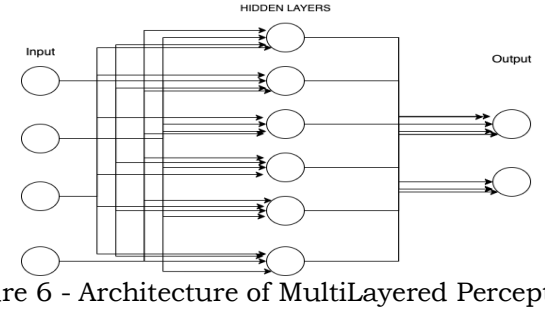

Figure 6 - Architecture of MultiLayered Perceptron

Dataset

Dataset consists of 100 years of rainfall observations in westbengal.This is recorded and publicly available. Using this data directly gives less accurate results. So, pre-processing or refining data is mandatory before supplying the

data to a model. Rainfall data has a major effect in raising the water level of the place further, leading to flooding in the area even though the flood occurrence also depends on the altitude and sea level of the particular place. This paper considers rainfall as a major factor in the occurrence of floods. These rainfall values are pre-processed as per requirements of the individual model. Model's output is recorded as Flood Occurrence Rate (FOR) for each and every model. The Dataset consists of the Rainfall value and Month with the Year of the record. A record of 105 years (1901 - 2015) is used to predict the FOR[16].

As this paper has considered supervised algorithms for prediction, all the specified models are provided with the supervision values of flooding. A new column is generated using the average rainfall value. This new one tells the occurrence of the flood(Binary - Whether YES/NO).

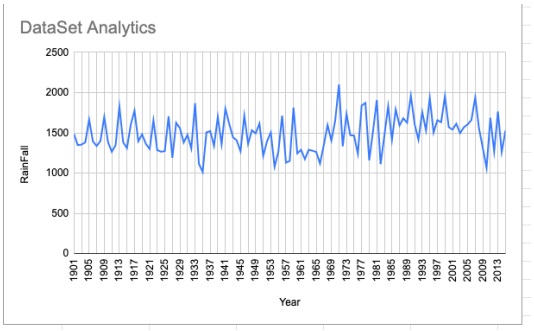

Figure 7 - Data set analytics

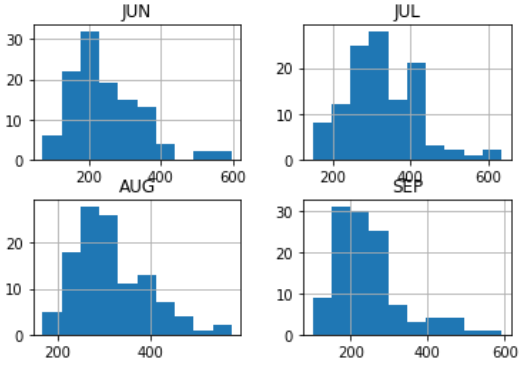

Figure 8 - History of the rainfall

Preprocessing

Dataset with the above-mentioned properties are submitted to analyze the data. Based on the model's preference, the preprocessing is done using different methods like Minmax scaler, which helps for scaling features, and model_selection for setting a blueprint to analyze data.

Results and Discussion

KNN classifier is a supervised algorithm, it is given with actual ground truth information. Neighbor module is used for more accuracy as it uses Ball Tree, KDtree for prediction. KNN-classifier determines whether the training data's FOR is true/false. As discussed earlier, a new column is generated with respect to the average value of Rainfall of all the records. It predicts the value as per the nearest neighbors. After completion of training, Model is tested with other data and then the accuracy of this particular model is given as 68.33%.
Logistic Regression predicts the binary output with previous observations or by recalling the recent observations. MinMax scaler is applied on the dataset for extracting and scaling the features upto the range. The Cross-Val-Score module is used for evaluating the scores.Accuracy score, recall score and roc score are 82.608696 ,66.666667 and 79.761905 respectively.

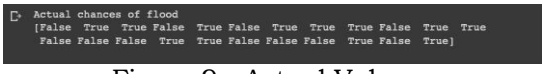

Figure 9 - Actual Values

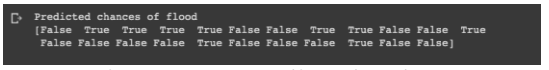

Figure 10 - Predicted Values

The Decision Tree Classifier helps in classifying the problem by the divide and conquer rule.Tree with binary classification at each level takes the pre-processed data directly. The Decision Tree classifier bisects the space on the graph into small regions for classifying the data. Accuracy score, recall score and roc score are 86.956522
,100.000000 and 89.285714 respectively.

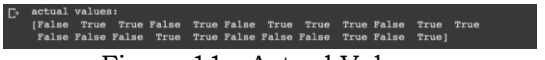

Figure 11 - Actual Values

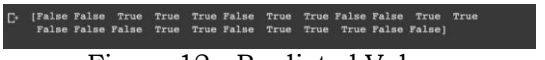

Figure 12 - Predicted Values

Random forest works similarly to decision trees. Random forest gives the output of multiple decision trees. Random forest is implemented here with a depth of 3. Random forest-produced accuracy score, recall score and roc score are 86.956522 ,77.777778 and 85.317460 respectively.

Ensemble learning uses the knowledge of multiple models or architectures to classify the test object. This Ensemble learning model uses all the above-mentioned machine learning algorithms i.e.., KNN, Logistic Regression, Decision tree classifier and Random Forest Classifier. This one has predicted with test accuracy 81.20%.

Long-short term memory Modelis a recurrent neural network. So, this is given by extending the dataset with some precision values. Dataset size is increased with negligible size precision values for better plotting and other result views[9].

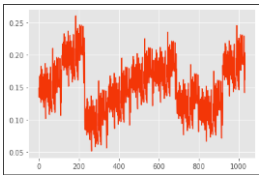

Figure 13 - Dataset after presicional

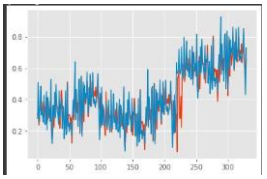

Figure 14 - actual vs prediction increase plot

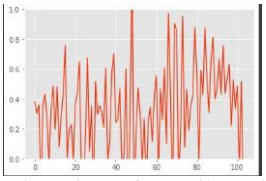

Figure 15 - Plot of Predicted Values

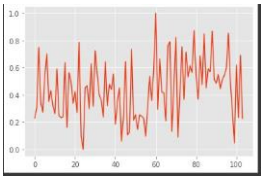

Figure 16 - plot of actual values

Dataset after precision is shown in Figure 13. MinMax scaler and other modules are also used for extracting the feature. Precision values train and test values are shown in Figure 14. When new test values are given to the trained model, it produces the results as per the plot shown in Figure 15 whereas ground truth or actual values are in Figure 16. Final accuracy of the model is approximately 75%.

Multi-Layered Perceptron is also a Deep Learning model which might require a high volume of data for good predictions. So, this model is also supplied with precisioned data of the original dataset. As this is one of the neural networks which remembers large amounts of sequential data, it is good at remembering patterns. Figure 17 represents the Epochs vs Cost plot. This model gives an accuracy of 88%.

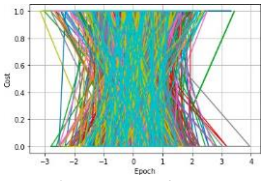

Figure 17 - Epochs vs Cost

S.No	Name of the Model	Machine Learning / Deep Learning	Accuracy
1	K Nearest Neighbour	Machine Learning	68.33%
2	Logistic Regression	Machine Learning	82.60%
3	Decision Tree Classifier	Machine Learning	86.95%
4	Random Forest Classifier	Machine Learning	86.95%
5	Ensemble Learning	Machine Learning	81.20%
6	Long-Short Term Memory	Deep Learning	75%
7	Multi-Layered Perceptron	Deep Learning	88%

Figure 18 - Accuracy results of all models

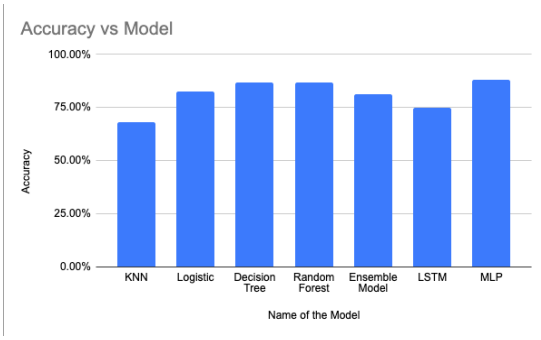

Figure 19 - Accuracy vs Model

Table 1 shows all accuracies produced for the supplied rainfall data.The same is shown in Figure 19.

Conclusion

Flood Prediction is calculated using different Machine Learning Models (KNN, Logistic Regression, Decision Tree Classifier, Random Forest Classifier), Ensemble Learning Model of Machine Learning models and Deep Learning Models (LSTM, MLP classifier). The highest accuracy of 88% is generated by Multilayered Perceptron. Being a Deep Learning Model, MLP can learn non-linear data in real time which results in better accuracy compared with other models. In future, our model will be implemented in graphics processing units to get high throughput along with faster performance.

References

1. Fang, Zhice, et al. "Predicting flood susceptibility using LSTM neural networks." Journal of Hydrology 594 (2021): 125734.
2. Haribabu, S., et al. "Prediction of Flood by Rainfall Using MLP Classifier of Neural Network Model." 2021 6th International Conference on Communication and Electronics Systems (ICCES). IEEE, 2021 Wright, Raymond E. "Logistic regression." (1995).
3. Sankaranarayanan, Suresh, et al. "Flood prediction based on weather parameters using deep learning." Journal of Water and Climate Change 11.4 (2020): 1766-1783.
4. Tanim, A.H.; McRae, C.B.;Tavakol-Davani, H.; Goharian, E.Flood Detection in Urban AreasUsing Satellite Imagery and Machine Learning. Water 2022,

14, 1140. https://doi.org/10.3390/w14071140
5. I. R. Widiasari, L. E. Nugroho and Widyawan, "Deep learning multilayer perceptron (MLP) for flood prediction model using wireless sensor network based hydrology time series data mining," 2017 International Conference on Innovative and Creative Information Technology (ICITech), 2017, pp. 1-5, doi: 10.1109/INNOCIT.2017.8319150.
6. Rahimzad, M., Moghaddam Nia, A., Zolfonoon, H. et al. Performance Comparison of an LSTM-based Deep Learning Model versus Conventional Machine Learning Algorithms for Streamflow Forecasting. Water Resour Manage 35, 4167–4187 (2021).
7. https://doi.org/10.1007/s11269-021-02937-w
8. Hu, Caihong, et al. “Deep Learning with a Long Short-Term Memory Networks Approach for Rainfall-Runoff Simulation.” Water, vol. 10, no. 11, Oct. 2018, p. 1543. Crossref, https://doi.org/10.3390/w10111543.
9. https://machinelearningmastery.com/gentle-introduction-l ong-short-term-memory-networks-experts/
10. Zhou, Q., Teng, S., Liao, X., Situ, Z., Feng, J., and Chen, G.: A deep learning technique-based data-driven model for accurate and rapid flood prediction, Hydrol. Earth Syst. Sci. Discuss. [preprint], https://doi.org/10.5194/hess-2021-596, in review, 2022.
11. https://www.analyticsvidhya.com/blog/2021/06/understa nding-random-forest/
12. Collier, C. G. "Flash flood forecasting: What are the limits of predictability?." The Quarterly Journal of the Royal Meteorological Society 133.622 (2007): 3-23.
13. Zhang, Xiaoqi, et al. "Real-time reservoir flood control operation for cascade reservoirs using a two-stage flood risk analysis method." *Journal of Hydrology* 577 (2019): 123954.
14. Gude, Vinayaka, et al. “Flood Prediction and Uncertainty Estimation Using Deep Learning.” *Water*, vol. 12, no. 3, Mar. 2020, p. 884. *Crossref*, https://doi.org/10.3390/w12030884.
15. Hou, Jingming, et al. "Rapid forecasting of urban flood inundation using multiple machine learning models." *Natural Hazards* 108.2 (2021): 2335-2356.https://www.kaggle.com/rajanand/rainfall-in-india?selec t=rainfall+in+india+1901