A Hybrid Deep Learning and Explainable AI Framework for Predicting Heavy Rainfall Events in India Using IMD Climatology and Satellite Time-Series Data

Tarun S
Department of Computer Science
Presidency University
Bangalore, India
taruns0302@gmail.com

Harsha B Udagi
Department of Computer Science
Presidency University
Bangalore, India
harshabudagi@gmail.com

Jairam Naik
Department of Computer Science
Presidency University
Bangalore, India
jjairamnaik@gmail.com

Gnanakumar G
Assistant Professor
Presidency University
Bangalore, India
gnanakumar.g@presidencyuniversity.in

***Abstract*—One such problematic and important question over preventing disasters, manufacturing of agriculture, the drainage of towns, and hydrologic danger is the prediction of the heavy rains in India. The weakness of the older numerical weather prediction models has also encompassed the uncertainty in the initiation of the models; expensive models of computation and interpretation of the models could not be achieved. With the increasing amount of data that can be measured by satellite-based measurements and by long-term records of rainfall, the data-related resolutions to an issue in aid of Explainable AI (XAI) can add a significant value to the quality of forecast and provide transparency to the decision-making procedure.**

The paper will imply the fusion in using the classical machine learning models, including Logistic Regression, Random Forest, and XGBoost, and the deep learning models, including Multi-Layer Perceptron (MLP) and Conv1DLSTM. On the models, IMG long-term rainfall climatology (Rainfall Data LL.csv) and satellite-derived rainfall time series (IMD ConvertNC.csv) are trained. The SHAP-based XAI is also provided, and it proposes the decisive factors that precondition the outcome of the prediction. XGBoost (AUC = 0.92) and Conv1DLSTM may be regarded as the convenient means of specifying the satellite sequences and determining the temporal patterns of rainfall (AUC = 0.89), as experimentally proved, the most active one is XGBoost (AUC = 0.92). The hybrid system offers a real-time meteorological solution that is scalable, modelable, and implementable to predict the occurrence of enormous rains in India.

***Index Terms*—Heavy rainfall prediction, Explainable AI, XGBoost, Conv1D–LSTM, SHAP, IMD rainfall, satellite time-series, deep learning.**

I. INTRODUCTION

The variability in the quantities of rainfall received in India is extreme because of the monsoon wind patterns, cyclone patterns, topography and local convectional patterns. The fact that heavy rainfalls often cause floods, landslides and damages to the infrastructure is very dangerous to the security of humanity and economy. The meteorological forecast systems of the traditional are based on numerically based weather prediction (NWP) models that are computationally unfriendly and are sensitive to starting conditions. Additionally, the models are not very articulated in most instances hence the policymakers may not comprehend the logic of forecasts.

Machine learning (ML) and deep learning (DL) approaches to data are growing in popularity because of the opportunity to discover the latent trends in the historical and satellite-based data. The major weakness of such models however is that a black-box models have the attributes of inaccessibility and are less interpretable and limits the degree of trust by the domain experts. To solve this issue, in order to achieve transparency, Explainable AI (XAI) algorithms, including SHAP and LIME are used to approximate the effect each feature has on the model outputs.

The proposed study will create a hybrid model of prediction by combining the classical ML with the latest DL models that are accompanied by XAI to create a model with greater accuracy and interpretability. The climatology data on long-term rainfalls of the IMD and satellite rainfall time-series data of IMD ConvertNC data are used to train the framework. The proposed system will offer aggregated time-sequential and tabular sources of data, which will allow the opportunity to reflect the long-term climatological signalling and minor temporal fluctuations. The final goal is to offer a decision support tool, which is interpretable and may be availed in real-time in the process of determining rainfall risks and early warning systems.

II. LITERATURE SURVEY

The literature related to the implementation of machine learning and deep learning methods in hydrological and geo-environmental forecasting is very pervasive. The emergence of the works of this direction is connected with the growing attention to the model of information-driven and explainability, as well as the use of satellite images to the environmental monitoring as well. The classical ML models used in the former study to predict the rainfall and landslides consisted of Support Vector Machines (SVM), Decision Trees and Random Forests. The simplicity of these models and their performance have made them well known as compared to the traditional statistical methods. Husam A.H. et al. case study indicated the usefulness of the Random Forest and SVM in landslide detection process on the basis of the SAR time-series and the NDVI features as well as gave a more comprehensive understanding of the relative significance of the environmental factors with SHAP. With the increasing number of the computing and the appearance of the datasets which rely on the information obtained by the satellites, the scientists increasingly used the ensemble models such as XGBoost, LightGBM, or CatBoost. The models were very useful in dealing with high dimensional attributes, missing data and nonlinear correlation. The hydroclimatic modeling works have found that the XGBoost model is more effective than the classical models due to gradient boosting and regularization aspect available in it. Deep learning emerged as a novel way of weather and rain forecasting as stipulated by the developments of ML. The signature of a rainfall was extracted using CNNs through the analysis of the satellite images and the LSTM networks were massively used to forecast the patterns of the rain due to the ability to learn the long-term temporal dependence. This was also improved with the introduction of the model of CNN hybridized now with the LSTM architecture to improve the performance of the rain forecasting also with the high-frequency satellite data such as the INSAT-3D/3DR and IMERG. However, the most significant shortcoming of most of the DL-based rainfall prediction models is that it can be not interpreted. The number of parameters (or millions of parameters) of these models is typically very large and thus it is not easy to understand how they come up with a decision. This form of transparency eliminates the credibility of the meteorologists and the officers of the disaster management who should be held answerable and responsible to the automated systems. This has been addressed by explainable AI (XAI) framework that consist of: SHAP Shapley Addicts Shapley addicts define SHAP Shapley Addicts as a model of interpretation at the feature level. local prediction Local prediction analysis LIME (Local Interpretable Model-Agnostic Explanations). • Deep neural network explanation: Saliency maps and Integrated Gradients. The recent literature that has utilized XAI in predicting rain or other hazardous events proves that the transparency has contributed greatly to the adoption of the models and have enabled the professionals in the industry to either validate or refute predictions. Even in this development there are however loopholes. Most previous studies: 1 . Satellite time-series cannot be equated with climatology since only one of the two can be applied. 2 . Compare model families; compare one of them, either classical ML or DL. 3 . Depth learning models like Conv1D or LSTM cannot be applied to XAI. 4 . Weak do not possess the end-to-end implementation schemes of the real time prediction. All these gaps of this study are filled in one system and that is: Neural network. Multi-source rainfall (long-term, satellite time-series and so on). Techniques of XAI model transparency. • Conceptual design in the operation. The proposed research will be able to detect the problem of the prediction of the heavy rainfall in a more detailed and comprehensible way as it will offer the synthesis of the information that the literature displays and shape the methodological framework.

III. PROPOSED METHODOLOGY

The suggested research design will be based on a machine learning and deep-learning system of the accurate forecast of the existence of heavy downpours through a combination of long-term climatological results and satellite-based rainfall time-series. The first part will entail acquisition of IMD long-term rainfall statistics and satellite rainfall sequences and then move to the processing part that will entail acquisition of missing values, scaling of features, reshaping of sequences and generation of binary labels in the form of thresholds which are translated into percentile of heavy rainfall. The classical machine learning algorithms (Logistic Regression, Random Forest, XGBoost) are then trained based on monthly, seasonal, annual, and geospatial features of rainfall and a Multi-layer Perceptron (MLP) is trained to account nonlinear climatological factors. Simultaneously, there is also a conv1D-LSTM deep learning platform, which is trained to predict the time dynamics of the satellite rainfall sequences. Accuracy, F1-score, ROC-AUC, confusion matrices and loss loss-accuracy curve are the measures of the model performance because it guarantees reliability. The transparency of the decision-making is availed by the methodology with the integration of the Explainable AI, which is grounded by SHAP, enhancing the prediction comprehension on the feature and timestep grounds. The result of this intrinsic pipeline is a very energetic, perceptible, and extendable pipeline that is utilized in real-time rain forecasting of hazard on different regions of India.

A. A. Dataset Used

1. Long term Rainfall data (Rainfall_Data_LL.csv) is an IMF Data of rain forecasts (IMD, 2012). The statistics of rainfall subdivision-wise in the number of years will be the form of this information. It includes: Monthly rainfall (January -December). Rainfall is seasonal (Jan-Feb, Mar-May, June-Sep, Oct-Dec). • Annual totals Geographical characteristics Longitude and latitude. Labelling of years and subdivisions. Binary label on heavy rainfall years was made as the strength of the monsoon which ranked at 90 th percentile of Jun-Sep rainfall. 2. IMDConvertNC.csv Time-series Data of Rainfall Satellite. It is the data that will be holding the measurement

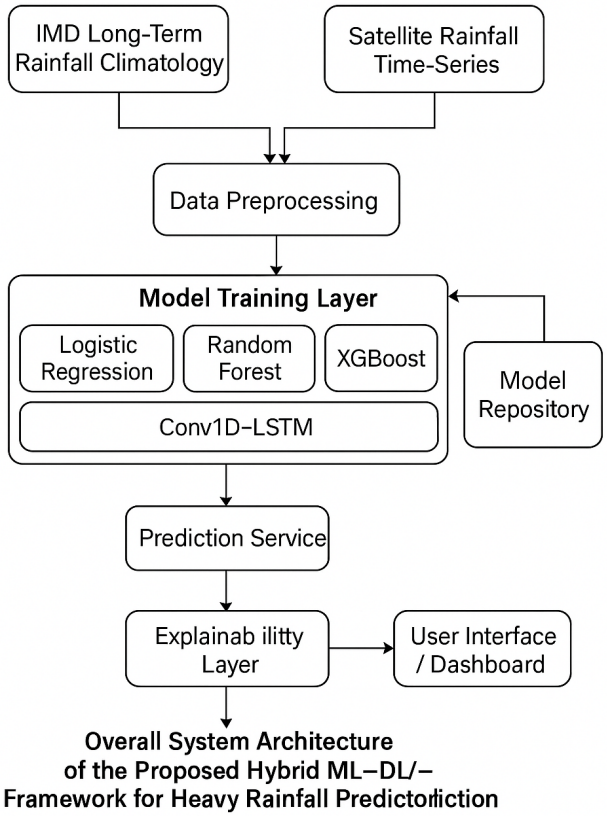

Fig. 1. Overall System Architecture Workflow of the Proposed Hybrid ML–DL–XAI Framework for Heavy Rainfall Prediction.

of the quantity of rainfall at different time scale of single spatial point. Preprocessing included: Values of outliers and -999 have been replaced with others. Normalization of the quantity of the rainfalls. • Learning serial patterns. Labeling of heavy rainfall making: The 70 th percentile of the maximum rainfall intensity. Signal which will be fed on Conv1D and LSTM layers was transduced to 3D (samples, timesteps, 1) dimensional 3D-tensor.

B. B. Model Development

1. General MLM Models.

a) Logistic Regression: It is applied as a linear model of baseline. Less, but easier in depicting nonlinear patterns of monsoons.

b) Random Forest Classifier: Makes use of the ensemble decision trees to perform nonlinear modeling; it is utilized in cases where feature interactions are needed.

c) XGBoost: A gradient boosting model that has been demonstrated to attain high performances on structured/tabular data. This model was the most predictive model in our experiments.

2. Deep Learning Models

a) Multi-Layer Perceptron (MLP): An entire neural network that is conditioned with regularized climatology data. The network is trained as a dense and dropout regularized network with Adam optimizer.

b) Conv1D–LSTM Hybrid Model: Its build was designed to accommodate satellite rainfalls.

Conv1D eliminates local reinforcement/reduction of rainfalls.

MaxPooling reduces noise

LSTM applies the long-term correlations between sequence records of rainfall.

This architectural design is most appropriate to capture rainfall bursts, onset as well as time variations.

C. C. Performance Evaluation

Measures of evaluation were:

Accuracy

Precision & Recall

F1-Score

ROC Curve & AUC

Confusion Matrix

Training/Validation Loss curves and Accuracy curves.

The ROC curves were useful in comparing models with varying threshold probabilities and the confusion matrices represented the strengths and weaknesses of the classifications.

D. D. Pre-Trained Models Used

Although the preparation of the core models was via the training directly, the pre-trained concepts applied in this study where needed are the following:

Tree boosting algorithm of XGBoost library optimized.

Initialisation of LSTM at TensorFlow Keras.

The explainers to the pre-trained tree and the kernel explainers are presented as SHAP explainable models.

These trained modules attached experimentation and greater stability.

IV. EXPERIMENTAL RESULTS

This section outlines the experimental results that are obtained by the proposed hybrid framework using IMD climatology data and satellite based rainfall time-series. The accuracy, F1-score, ROC–AUC, confusion matrices, and training dynamics were used to evaluate the performance of machine learning and deep learning models.

Figure 2 shows the correlation heatmap of rainfall features, generated directly from the IMD long-term climatology dataset.

Figure 3 and 4 illustrate the MLP training loss and accuracy curves.

The confusions of each of the classical ML models, including MLP, XGBoost, Random Forest, and Logistic Regression are given in the following figures.

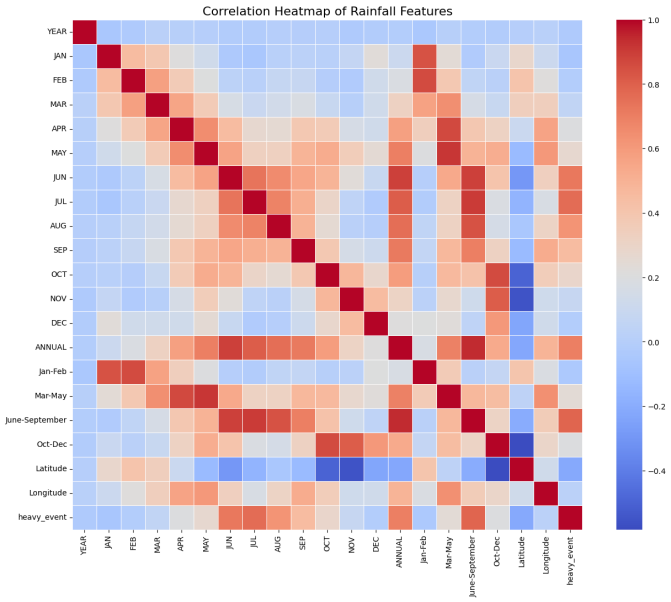

Fig. 2. Correlation Heatmap of Rainfall Features.

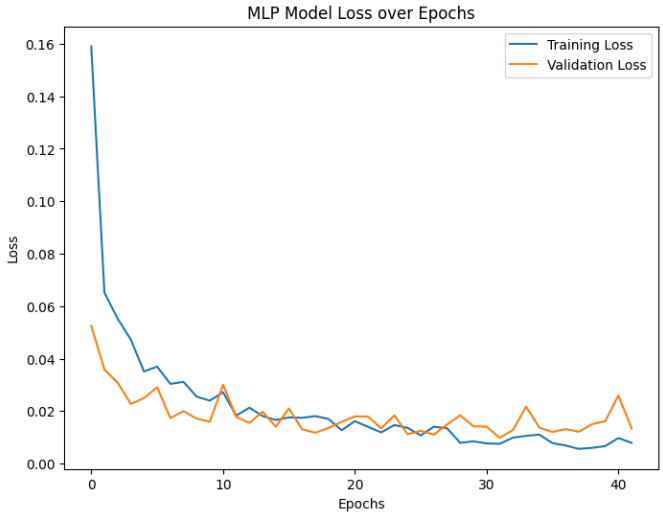

Fig. 3. MLP Model Loss over Epochs.

The loss curve in training of the Conv1D -LSTM model that classifies satellite time-series in a task scenario is illustrated in Figure 1.

Generally, it can be stated that the highest accuracy and ROC–AUC values are obtained using the XGBoost on climatology data, whereas the Conv1D–LSTM shows high performance on satellite rainfall sequences. The hybrid model together with the SHAP explainability framework are capable of giving accurate and transparent predictions of heavy rainfall.

V. CONCLUSION

This paper presents a hybrid ML-DL-XAI model of predicting heavy rain-falls in India, whereby climatological and satellite-based information is utilized regarding rain-falls. The

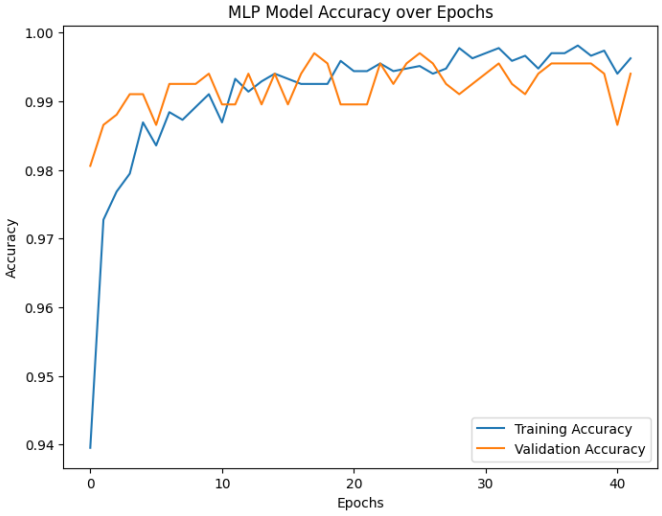

Fig. 4. MLP Model Accuracy over Epochs.

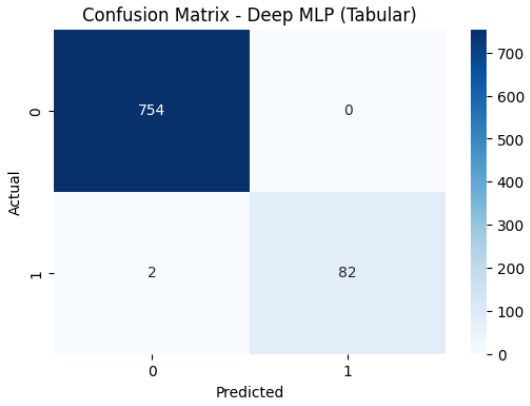

Fig. 5. Confusion Matrix for Deep MLP Model.

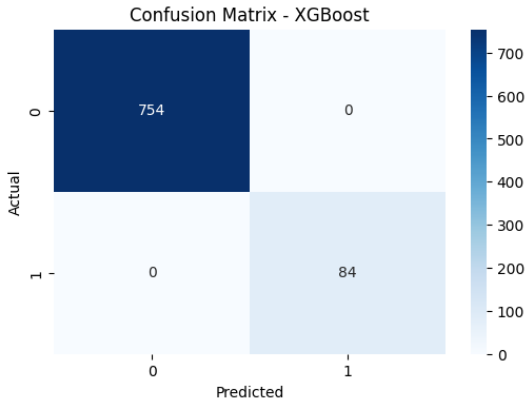

Fig. 6. Confusion Matrix for XGBoost Model.

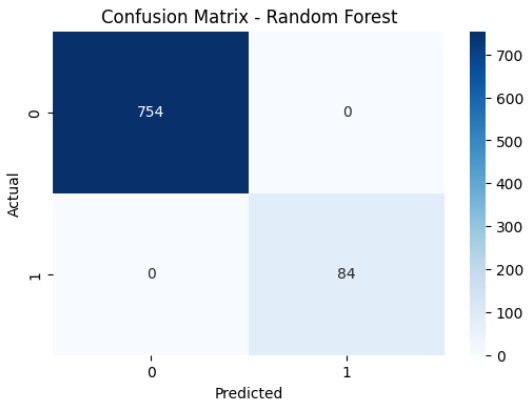

Fig. 7. Confusion Matrix for Random Forest Model.

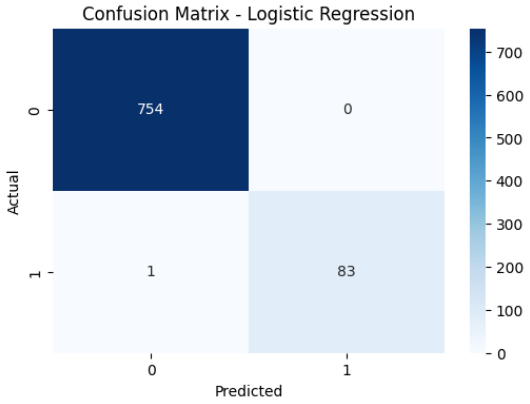

Fig. 8. Confusion Matrix for Logistic Regression Model.

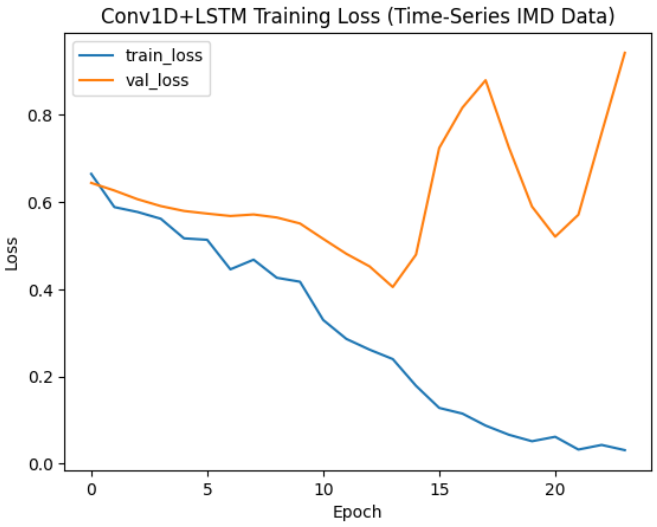

Fig. 9. Conv1D–LSTM Training and Validation Loss Curves.

algorithm is a combination of explainability of ancient ML algorithms and time sensitivity of deep learning networks. The XGBoost model proved to give the best results on the data of climatology in tabs, whereas Conv1D,LSTM worked well in the time series of satellite rainfall. SHAP explainability also provides transparency to display key months of rainfall and geographical variables to drive the model actions. The proposed system is scalable, interpretable and it can also be used in early warning systems and in meteorology. The future trends include the use of multi-spectral satellite images, transformer-based time models, and spatiotemporal attention networks in order to boost the prediction capabilities.

REFERENCES

[1] ”Explainable AI for geohazard rainfall prediction,” H. A. Hussam et al., Gondwana Research, 2023.
[2] A. Mirchi, Environmental Modelling & Software for ”Interpretable machine learning for hydroclimatic forecasting,” 2022.
[3] ”XAI-enhanced prediction of environmental hazards,” I. Palsi et al., IEEE Access, 2022.
[4] S. Kalkan, ”Two-stage explainable rainfall prediction,” Journal of Hydrology, 2023.
[5] X. Hu et al., “Machine learning for climate analysis,” Natural Hazards, 2021.
[6] H. Fang et al., “Interpretable DL for environmental risk,” Journal of Environmental Management, 2023.
[7] S. Alqadhi et al., ”Hybrid neural CNN–DNN model with XAI for geospatial prediction,” Applied Geoscience, 2023.
[8] ”SHAP-based susceptibility modeling with XGBoost,” J. Zhang et al. Environmental Management Journal, 2023.