IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

Explainable Deep Learning for Rainfall Prediction: A CNN-XGBoost Hybrid Approach in Northern Bangladesh

Md Safayet Islam[1], Md Shafiuzzaman[1, 2], Golam Mahmud[1], Nabila Nowshin[1], Parisa Reza[1], Jahid Hasan[4], Md Nahiduzzaman[1,3], Mohamed Arselene Ayari[5], Amith Khandakar[6]

[1]Department of Electrical and Computer Engineering, Rajshahi University of Engineering and Technology, Rajshahi, Bangladesh.
[2]Department of Computer Science & Engineering, North Bengal International University, Rajshahi, Bangladesh.
[3]Department of Compute Science, Royal Melbourne Institute of Technology, Melbourne, Australia.
[4]Department of Urban and Regional Planning, Rajshahi University of Engineering and Technology, Rajshahi, Bangladesh.
[5]Department of Civil and Environmental Engineering, Qatar University, Doha 2713, Qatar.
[6]Department of Electrical Engineering, Qatar University, Doha 2713, Qatar.

Corresponding author: Mohamed Arselene Ayari (e-mail: arslana@qu.edu.qa).

ABSTRACT Accurate precipitation forecasting is crucial for evaluating various hydrological processes. This research explores the application of deep learning models for rainfall prediction in Northern Bangladesh, focusing on the comparative performance of five models: CNN, LSTM, XGB, Transformer-XGB, and CNN-XGB. Two distinct datasets were utilized to assess the effectiveness of these models. Among them, the CNN-XGB hybrid model consistently demonstrated superior performance across all evaluation metrics, establishing it as the most reliable predictor in this study. To enhance the interpretability of the proposed CNN-XGB model, we deployed the SHAP (SHapley Additive exPlanations) explainer, providing insights into the model's decision-making process. This research underscores the potential of hybrid models in improving rainfall prediction accuracy while offering transparency through explainable AI techniques.

INDEX TERMS CNN, hybrid model, LSTM, machine learning, Northern Bangladesh, rainfall forecasting, XGBoost

I. INTRODUCTION

A country like Bangladesh, located in the tropical region, experiences diverse and dynamic rainfall patterns, which are critical for its agrarian economy and overall socio-economic landscape. As rainfall is the most important meteorological parameter that impacts human activities at diverse scales, including agriculture, water resource management, tourism, urban construction, hydroelectric power plants, and so forth its prediction is highly significant. Accurate early rainfall predictions are essential for managing water-related disasters like floods, droughts, and storms for countries like Bangladesh thriving on agro-based economies [1],[2]. Since Bangladesh is prone to natural disasters, such as floods and cyclones, due to its geographical location, accurate rainfall predictions play a pivotal role in disaster preparedness. These forecasts enable the nation to anticipate potential flooding events, facilitating timely evacuations and other essential disaster management measures [3]. predicting how geophysical systems respond is incredibly intricate due to the multitude of causal factors from both the oceanic and atmospheric realms and the complex, nonlinear relationships between them. Among the most challenging tasks for modelers is accurately forecasting rainfall across various time scales, be it daily, monthly, seasonal, or annual for effective risk management and disaster preparedness [4].

In the recent past, researchers have devised numerous methods for predicting rainfall. These forecasting models typically fall into two main categories: physical models and data-driven models. The physical model utilizes statistical methods to analyze the relationship between rainfall and various geographic and atmospheric factors, including geographic coordinates like latitude and longitude, as well as atmospheric variables such as pressure, temperature, wind speed, and humidity. However, the complex nature of rainfall, characterized by its nonlinearity, poses a significant challenge to accurate prediction using these methods [5] Thence, Artificial Neural Networks (ANNs) have emerged as the favored data-driven approach for forecasting dynamic, nonlinear, and non-stationary hydrological variables as computationally less expensive forecasting approaches. As the volume of data grows, Artificial Neural Networks (ANNs) struggle to achieve improved performance. In such scenarios,

the advantage of deep learning (DL) techniques becomes apparent. DL methods can leverage larger datasets to extract more profound insights, making them well-suited for modeling complex hydrological variables. Thus, DL algorithms are increasingly gaining favor among researchers for their ability to handle and interpret extensive predictor-predicted databases [6].

Long Short-Term Memory (LSTM) networks are a type of recurrent neural network (RNN) architecture specifically designed to address the limitations of traditional RNNs in capturing long-term dependencies in sequential data. As it works through feedback connection, it shows massive efficiency in sequence prediction, especially for the time series data. LSTM stands out as the predominant deep-learning method employed in hydrological forecasts with the ability to handle large datasets with diverse complexities [7].

Many researchers have created hybrid models by merging Empirical Mode Decomposition (EMD) or its variations with different methodologies, including linear regression and machine learning, for hydrological forecasting. Typically, in these hybrid approaches, the time series data of the variable are broken down into components. Then, these individual components are forecasted separately using either linear or non-linear regression techniques. These predicted components are combined to generate the overall forecast [8] Methods like Empirical Mode Decomposition (EMD) are considered highly suitable for decomposing non-stationary and nonlinear data due to their adaptability to the data's characteristics for time series signals.

To forecast monthly rainfall, Johny et al. (2021) proposed a deep learning approach, a novel hybrid modeling framework incorporating LSTM and MEMD, based on the TDICC analysis algorithm. This combination led to a significant reduction in modeling complexity and swifter execution [4]. An AI-based framework for debris flow rainfall prediction was proposed by Zhao et al. (2022). It could automatically estimate the probability that any particular rainfall event would cause a debris flow by using data from 367 rainfall events in central China. ETs model achieved the highest possible efficiency [9] Animas et al. (2022) offered a study to show an ensemble model developed via an Automated Machine Learning approach and an XGBoost baseline model were compared with LSTM-Networks-based models, Stacked-LSTM-based models, and Bidirectional-LSTM Networks model. Models typically overfitted training data and are unable to come up with precise predictions in test and validation sets[5].The outcomes of regression techniques proposed by Mahabub et al. (2019) are compared to the forecast-based models that are currently in use, demonstrating the enhanced efficacy with a modest degree of error of machine learning-based models over conventional methods[10].Tae-wong et al. developed the Space-Time model, which covered the brief-term temporal and spatial factors of the daily rainfall event[11].A number of regression algorithms employed by Zaman et al. revealed their comparative performance in their work on an ML-based rainfall anticipation system for Bangladesh[12].The ability of ANNs to effectively deal with non-linearity in rainfall data and the fact that they require little to no comprehension of the relationships between the variables under consideration are significant factors that have helped the expansion of their use as a method for rainfall anticipating (Liu et al., 2019)[13]. Non-linear ANN models outperformed MLR models in Western Australia spring rainfall estimation with respect to statistical errors and Pearson correlation. (Hossain et al. 2019)[14]For monthly streamflow and rainfall forecasting, Ni et al. (2020) proposed one WLSTM model employing a trous algorithm of wavelet transform was used to carry out series decomposition, and a coupled convolutional neural network model called CLSTM was used to extract temporal features[15].Poornima et al. (2019) suggested deploying a Recurrent Neural Network (RNN) based on Intensified Long Short-Term Memory (Intensified LSTM) to forecast rainfall which was later contrasted with alternative models to show the improvement[16].A feed-forward back-propagation neural network algorithm has been used for the nonlinear nature of Indian summer monsoon rainfall forecasting where five neural network architectures using three layers of neurons with one input layer, one hidden layer, and one output layer) have been proposed (Singh et al. 2013)[17]. A model based on Ensemble Empirical Mode Decomposition (EEMD) has been suggested by Xiang et al. (2018) in which Support Vector Regression (SVR) is used for predicting short-period components, and Artificial Neural Networks (ANN) are used for predicting long-period components [18]. For the aim of forecasting India's monsoon rainfall, a multivariate extension of EMD and Stepwise Linear Regression has been offered (Adarsh et al. 2017). In ISMR predictions, the recommended MEMD-SLR method has proven a definite advantage over the IMD operational forecast, M5 Model Tree, and multiple linear regression methods[19].Adaptive Ensemble Empirical Mode Decomposition-Artificial Neural Network (AEEMD-ANN) model is a particular form of adaptive hybrid modeling framework that anticipates ISMR by performing forecasts adaptively based on the inclusion of new information (Jhony et al. 2020)[20]The complicated relationship between the causal variables and the daily fluctuation of rainfall is better captured by a hybrid deep learning tactic which incorporates a one-dimensional Convolutional Neural Network (Conv1D) with an MLP model (Khan et al. 2020)[21].Three meteorological variables—temperature, dew point, and humidity—were used in Salman et al.'s (2018) hybrid approach, which combined an ARIMA model with DL-based LSTM[22] .A neural network-based approach for atmospheric condition predictions was proposed by Sawale et al. The authors employed a hybrid architecture that integrated the use of Back Propagation Network (BPN) and Hopfield Network (HN) techniques[23].Results of a proposed approach by Wu et al. which incorporates single spectrum analysis (SSA) and

modular artificial neural network (MANN) reveal that MANN has a number of notable advantages over existing models, most notably in daily rainfall predictions[24].The effectiveness of three machine learning and deep learning techniques for rainfall forecasting was compared by Aderyani et al. (2022) including a hybrid neural network (CNN), long-short-term memory, and optimal support vector regression. The outcome showed that the PSO-SVR and LSTM methods outperformed CNN by about the samemargin[25].For rainfall prediction in weather derivatives, a thorough assessment of the predictive performance of seven machine-learning techniques was conducted by Cramer et al. (2017). These included the state-of-the-art and six other acknowledged machine learning algorithms: genetic programming, support vector regression, radial basis neural networks, M5 Rules, M5 Model trees, and k-nearest neighbors [26].

Despite significant advancements in the application of deep learning and machine learning models for rainfall prediction, there remains a noticeable gap in the exploration of hybrid models that combine the strengths of multiple algorithms. Previous studies have largely focused on individual models like LSTM, CNN, and XGBoost, often neglecting the potential performance improvements that could be achieved through hybrid approaches. Additionally, there has been limited emphasis on the interpretability of these models, with most research prioritizing accuracy over understanding the model's decision-making process. This gap is particularly evident in region-specific studies, where the unique climatic conditions of areas like Northern Bangladesh have not been adequately addressed in existing literature. Many studies rely on a single dataset or a narrow range of data sources, limiting the generalizability of the model's performance. There is a need for more studies that utilize diverse datasets to comprehensively evaluate the robustness of rainfall prediction models.

This research bridges the identified gaps by introducing and thoroughly evaluating a CNN-XGB hybrid model for rainfall prediction in Northern Bangladesh. The study not only demonstrates the superior performance of this hybrid model across various evaluation metrics compared to individual models but also enhances model transparency by employing the SHAP explainer to interpret the model's decisions. By utilizing two distinct datasets, the research provides a robust comparative analysis and offers valuable insights into the applicability of advanced hybrid models in regional rainfall forecasting, particularly in areas with complex and dynamic climatic patterns like Northern Bangladesh.

II. DATASET

A. DATA DESCRIPTION

The first study area consists of different districts of the Rajshahi division located in the northern region of Bangladesh. The districts are: Joypurhat, Naogoan, Rajshahi, Chapainawabganj, Natore, Pabna, Sirajganj, and Bogura. Rajshahi division shares its border with the Indian states of West Bengal, to the west, Rangpur division, to the north, Dhaka division to the east, and Khulna division to the south. The daily rainfall data from January 2000 to December 2023 for the targeted area was downloaded from the website of NASA Power which is publicly available [27]. Descriptions of all parameters are shown in Table 1. The collected rainfall data by MERRA-2 was in mm/day.

TABLE 1-METEOROLOGICAL DATA PARAMETERS AND DESCRIPTIONS OF SATELLITE DATASET

Feature Name	Description
Year	Year (2000-2023)
DOY	Day of Year
PRECTOTCORR	Precipitation Corrected (mm/day)
T2M	Temperature at 2 Meters (C)
TS	Earth Skin Temperature (C)
T2M_RANGE	Temperature at 2 Meters Range (C)
T2M_MAX	Temperature at 2 Meters Maximum (C)
T2M_MIN	Temperature at 2 Meters Minimum (C)
QV2M	Specific Humidity at 2 Meters (g/kg)
RH2M	Relative Humidity at 2 Meters (%)
WS2M	Wind Speed at 2 Meters (m/s)
WS2M_MAX	Wind Speed at 2 Meters Maximum (m/s)
WS2M_MIN	Wind Speed at 2 Meters Minimum (m/s)
WD2M	Wind Direction at 2 Meters (Degrees)
WS10M	Wind Speed at 10 Meters (m/s)
WS10M_MAX	Wind Speed at 10 Meters Maximum (m/s)
WS10M_MIN	Wind Speed at 10 Meters Minimum (m/s)
WD10M	Wind Direction at 10 Meters (Degrees)

MERRA-2 (Modern-Era Retrospective Analysis for Research and Applications, Version 2) is a comprehensive atmospheric reanalysis dataset produced by NASA. It provides high-resolution climate data by adjusting a widespread range of observations from satellites, weather stations, and other sources into a climate model. From 1980 to the present, MERRA-2 provides comprehensive data on a range of atmospheric, land surface, and oceanic characteristics, including temperature, humidity, wind, and precipitation. This dataset is widely utilized in climate research, weather forecasting, and environmental studies because it helps scientists study past climate patterns, understand current weather, and predict future conditions. MERRA-2 is particularly valuable for studying long-term climate changes, extreme weather events, and the impacts of these changes on different regions and ecosystems.

Another monthly rainfall dataset was collected from the ground substation of Bogura, Dinajpur, Pabna, and Rajshahi which contained only rainfall (mm/month) from January 1950 to December 2019.

TABLE 2-FEATURE DESCRIPTION OF GROUND STATION DATASET

Feature Name	Description
Year	Year
Month	Month of a Year
Rainfall	Amount of rainfall (mm/month)

Overall, there is tropical monsoon climate with three distinct seasons in the Rajshahi division: first winter is between November and February, which is relatively cool with nearly no rainfall, summer which is between March and June, warm and characterized by thunderstorms, and monsoon is between July and October, with heavy rainfall.

The study area extends from 24° 21' N, 88° 30' E and 25° 02' N, 89° 30' E, and it is the most unconventional part of the country to suffer from unexpected rainfall events due to the existed climate pattern in the chosen cities. For example, Natore is considered the driest city, while Rajshahi is the city with the hottest summer days.

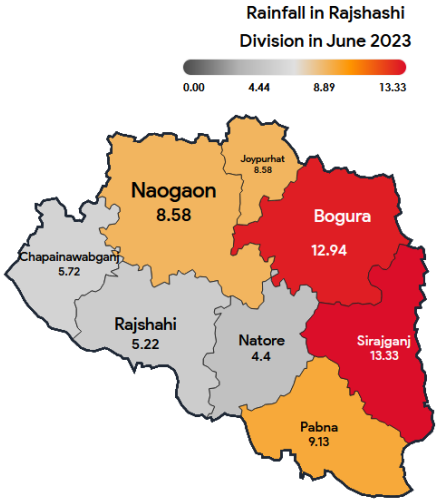

FIGURE 1. Average rainfall in Rajshahi division in June 2023.

Fig.1 illustrates the average rainfall in June 2023 in the targeted area. The highest rainfall is 13.33 mm/day in Sirajganj, and the lowest rainfall is 4.4 mm/day in Natore. Rainfall in Bogura is very close to the rainfall in Sirajganj due to the river Jamuna. And rainfall in Rajshahi, Chapainawabganj, and Natore are pretty close. Average rainfall in Naogoan and Joypurhat are equal. At last, the average rainfall in Pabna is quite different from its adjacent districts.

B. TOP POSITIVE CORRELATED FEATURE FOR RAINFALL PREDICTION

Table 3 depicts a detailed analysis to identify the top positive correlation among diverse meteorological parameters for rainfall prediction in different regions. The chosen feature comprises specific humidity at 2 meters above the ground (QV2M) which indicates the amount of water vapor near the surface, and relative humidity at 2 meters (RH2M) indicating the current amount of moisture in the air relative to its maximum capacity, wind speed at 2 meters (WS2M) reflecting how fast wind is moving near the ground, minimum temperature 2 meters (T2M_MIN), and wind speed at 10 meters. All data is formatted to two decimal places. The QV2M is the highest correlated feature with rainfall among all the parameters ranging between 0.44 and 0.48. Higher specific humidity indicates more moisture availability in the air, causing the formation of clouds and precipitation. Hence, RH2M appears as the second highest correlated parameter ranging between 0.38 and 0.41, indicating air is close to saturation and responsible for condensation and cloud formation. The correlations observed in the table are not particularly strong (i.e., not close to 1). To overcome this drawback feature engineering techniques will be applied subsequently to identify better relationships.

TABLE 3-TOP POSITIVE CORRELATION OF METEOROLOGICAL PARAMETERS WITH RAINFALL FOR SATELLITE DATA

District	QV2M	RH2M	WS2M	T2M_MIN	WS10M
Rajshahi	0.47	0.39	0.39	0.38	0.37
Bogura	0.46	0.4	0.41	0.39	0.39
Chapai	0.48	0.41	0.37	0.39	0.35
Joypurhat	0.47	0.38	0.37	0.40	0.34
Naogoan	0.47	0.38	0.36	0.39	0.34
Natore	0.46	0.40	0.39	0.38	0.37
Pabna	0.44	0.40	0.39	0.38	0.37
Sirajganj	0.45	0.39	0.4	0.38	0.38

C. TOP LEAST CORRELATED FEATURE FOR RAINFALL PREDICTION

In our experiment, we examined the least correlated meteorological parameters for rainfall predictions across diverse regions. Table 4 represents the data of the least correlated feature where each region is characterized by the mean diurnal temperature range (T2M_RANGE) which reflects the difference between daily maximum and minimum temperature, wind direction at 10 meters (WD10M), wind direction at 2 meters (WD2M), maximum temperature at 2 meters above the ground (T2M_MAX), and day of the year (DOY). The values of each feature are numeric and formatted to two decimal places. Among all the criteria T2M_RANGE shows the least correlation with rainfall across the district with

TABLE 4- TOP LEAST CORRELATION FEATURE OF METEOROLOGICAL PARAMETERS WITH RAINFALL FOR SATELLITE DATA

District	T2M_ RANGE	WD10M	WD2M	T2M_MAX	DOY

Rajshahi	-0.5	-0.25	-0.25	0.04	0.08
Bogura	-0.48	-0.21	-0.2	0.08	0.07
Chapai	-0.5	-0.27	-0.27	0.04	0.09
Joypurhat	-0.49	-0.22	-0.22	0.06	0.08
Naogoan	-0.49	-0.22	-0.22	0.06	0.08
Natore	-0.5	-0.25	-0.25	0.04	0.08
Pabna	-0.48	-0.26	-0.26	0.02	0.07
Sirajganj	-0.48	-0.25	-0.24	0.07	0.06

values ranging from -0.48 to -0.50. Hence, the variations in temperature range don't show a significant influence on rainfall predictions. Similarly, the wind directions at 2 meters and 10 meters above the ground also show the second highest least correlation with rainfall predictions. Since they are less influential, these meteorological features will be excluded from our considerations.

D. CORRELATION MATRIX

The correlation matrix in Fig.2 displays Pearson correlation coefficients between various variables, ranging from -1 to 1, indicating the strength and direction of their linear relationships. Red to dark red colors show high positive correlations, blue to dark blue show high negative correlations, and white to light colors indicate low correlations. Each variable is perfectly correlated with itself, as shown by the diagonal elements. Significant correlations include a very high positive correlation between T2M and TS (0.99), and T2M and T2M_MAX (0.94), while notable negative correlations include T2M_RANGE and T2M_MIN (-0.87). Moderate correlations are seen between QV2M and T2M (0.64), and WS10M_MAX and WS2M_MAX (0.71).

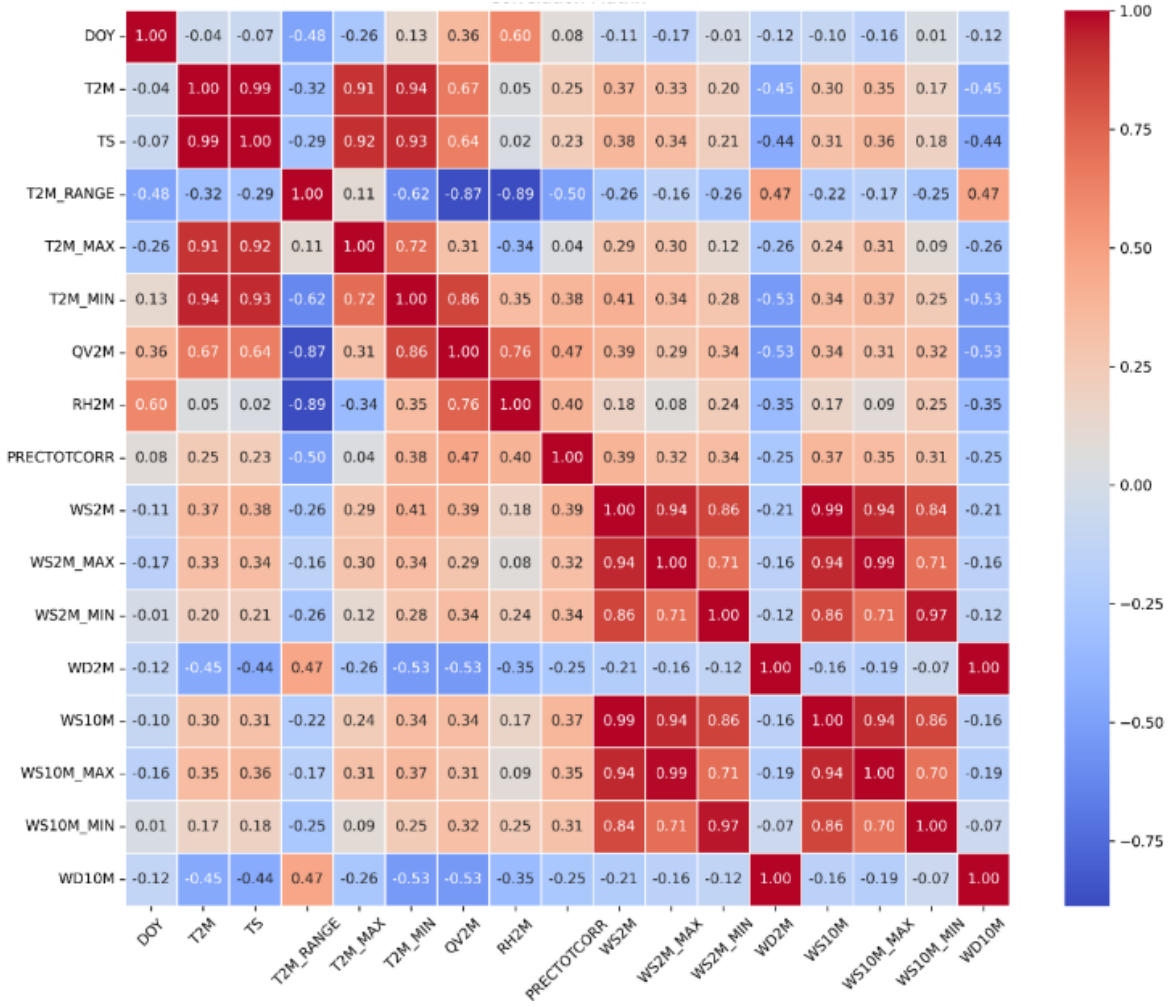

FIGURE 2 . **Correlation matrix of Rajshahi district**

Day of the year (DOY) shows low correlations with most variables. Temperature-related variables (T2M, T2M_MAX, T2M_MIN, T2M_RANGE) are highly correlated with each other, and wind-related variables (WS2M, WS2M_MAX, WS2M_MIN, WS10M, WS10M_MAX, WS10M_MIN, WD10M) also show moderate to high correlations. This matrix helps in understanding the interrelationships between

various climatic and meteorological variables, aiding in the development and refinement of predictive models.

The correlation matrix shows that PRECTOTCORR (Rainfall) has low positive correlations with temperature and humidity variables such as T2M (0.23) and QV2M (0.32), indicating slight increases in precipitation with higher temperatures and humidity. It has low negative correlations with wind speed variables like WS2M (-0.25) and WS2M_MIN (-0.26), suggesting a slight decrease in precipitation with higher wind speeds. Additionally, there is a moderate positive correlation with wind direction at 2 meters (WD2M, 0.39), indicating that certain wind directions are moderately associated with increased precipitation totals.

III. PREPROCESSING

A. FEATURE ENGINEERING

The feature engineering pipeline aims to enrich time-series data by extracting diverse characteristics that can enhance predictive modeling and analysis [28]. It begins by decomposing each time series into trend, seasonal, and residual components, enabling the capture of temporal patterns. Subsequently, autocorrelation and partial autocorrelation metrics are computed to quantify the data's self-similarity at different lags. Different statistical attributes such as lagged values, rolling means, and standard deviations are then calculated, providing insights into the data's dynamics and variability over time. Additionally, interaction features like the product of temperature and relative humidity, along with derived features such as temperature range, further augment the dataset, facilitating a more comprehensive understanding and utilization of the underlying temporal information.

The feature engineering proposed in this paper on both datasets for rainfall prediction enhances the original features by creating a variety of new ones. The main steps include seasonal decomposition, autocorrelation, partial autocorrelation, and various statistical measures which demonstrated in Fig. 3. Specifically, for each column, the procedure performs the following:

1. *Seasonal Decomposition:* This is done using periods of 2, 3, 5, 7, and 14. For each period, three new features are created: trend, seasonal, and residual components. This results in 15 new features per original column (3 features × 5 periods).
2. *Autocorrelation and Partial Autocorrelation:* For each of the 5 periods, autocorrelation (ACF) and partial autocorrelation (PACF) values are computed and stored as new features, adding 10 new features per column (2 features × 5 periods).
3. *Lag, Rolling Statistics, and Exponential Weighted Moving (EWM) Statistics:* For each of the 5 periods, the code calculates lagged values, rolling mean, rolling standard deviation, shifted standard deviation, EWM standard deviation, and EWM mean. This results in 30 new features per column (6 features × 5 periods).
4. *Interaction and Derived Features:* Two additional features are created based on interactions and derivations: the product of temperature and relative humidity (`T2M_RH2M`), and the temperature range (`TempRange`), defined as the difference between maximum and minimum temperatures.

In total, for each column, the feature engineering process generates 55 new features (15 from seasonal decomposition, 10 from ACF/PACF, and 30 from lag/rolling/EWM statistics). If the total number of columns are n, the total number of new features added to the dataset is (55 × n + 2). This extensive feature set aims to capture temporal patterns, correlations, and statistical properties to improve the predictive power of the rainfall prediction model.

The name of the generated feature should cover a general formula, *col_transformation_period*. Where *col* represents the original column name, *transformation* represents the type of transformation (e.g., trend, seasonal, residual, ACF, pacf, lag, rolling_mean, rolling_std, std, ewm_std, ewm_mean) and *period* represents the period used (e.g., 2, 3, 5, 7, 14). For example, `T2M_trend_3` represents the long-term trend component of the `T2M` temperature data, extracted using a seasonal decomposition with a period of 3-time step.

For the satellite dataset, the number of input features is 9 ('PRECTOTCORR', 'T2M', 'TS', 'QV2M', 'RH2M', 'WS2M', 'WD2M', 'WS10M', 'WD10M'). The feature engineering process generates 55 new features per input feature by applying transformations across 5 different periods (2, 3, 5, 7, and 14), resulting in (9 × 55 = 495) new features. Additionally, two more features through interaction and derived feature calculations are added. Including the 9 original features, the final dataset will have (9 + 495 + 2 = 506) features in total.

For the ground dataset, the input data contains only one variable (rainfall), and the feature engineering process generates 55 new columns. Since 11 new features are created for each period, and there are 5 periods, the total number of added columns is 55. Including the original rainfall column, the final data frame will have 56 columns in total.

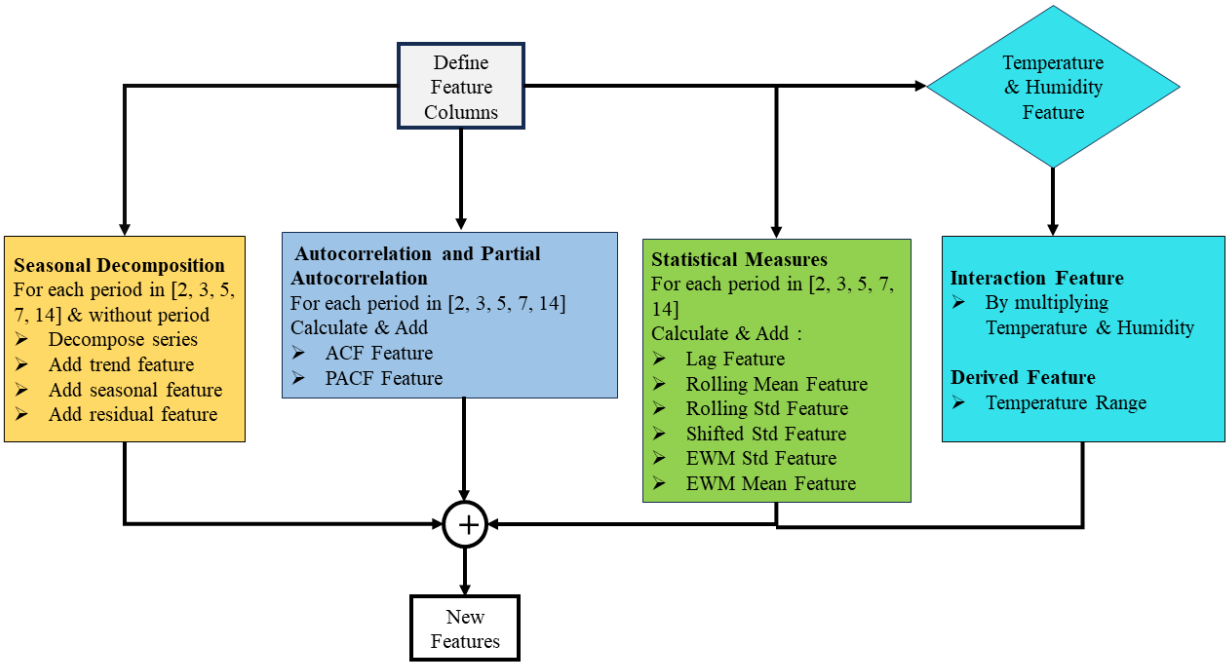

FIGURE 3. Flowchart of feature engineering.

TABLE 5- TOP 10CORRELATED FEATURES AFTER FEATURE ENGINEERING

District	Rajshahi	Bogura	Chapai	Joypurhat	Naogoan	Natore	Pabna	Sirajganj
PRECTOTCORR_trend_2	0.95	0.95	0.95	0.95	0.95	0.95	0.94	0.94
PRECTOTCORR_rolling_mean_2	0.9	0.91	0.9	0.9	0.9	0.9	0.9	0.89
PRECTOTCORR_trend_3	0.9	0.9	0.89	0.9	0.9	0.9	0.89	0.88
PRECTOTCORR_residual	0.86	0.86	0.86	0.85	0.85	0.86	0.86	0.85
PRECTOTCORR_ewm_mean_2	0.83	0.83	0.83	0.83	0.83	0.83	0.82	0.83
PRECTOTCORR_residual_14	0.82	0.81	0.81	0.81	0.81	0.82	0.82	0.81
PRECTOTCORR_trend_5	0.81	0.8	0.8	0.81	0.81	0.81	0.8	0.8
PRECTOTCORR_rolling_mean_3	0.8	0.8	0.8	0.8	0.8	0.8	0.79	0.8
PRECTOTCORR_ewm_mean_3	0.77	0.78	0.77	0.77	0.77	0.77	0.77	0.77
PRECTOTCORR_trend_7	0.74	0.74	0.74	0.74	0.74	0.74	0.72	0.73

B. RAIN CORRELATION TOP 10 AFTER FEATURE ENGINEERING

The analysis of the correlation values after the statistical feature engineering of various features with rainfall (PRECTOTCORR) across eight districts—Rajshahi, Bogura, Chapai, Joypurhat, Naogoan, Natore, Pabna, and Sirajganj is demonstrated in Table 5.

This analysis reveals that short-term trends and rolling means are the most highly correlated features, with values consistently around 0.94 to 0.95 for two-period trends and 0.89 to 0.91 for two-period rolling means. Three-period trends also show strong correlations, ranging from 0.88 to 0.90, highlighting the significance of short-term trend analysis in rainfall prediction. Residuals, which represent deviations from the trend, display correlations from 0.85 to 0.86, indicating their importance in capturing unexpected variations. Moderately correlated features include the exponentially weighted moving average over two periods (0.82 to 0.83) and fourteen-period residuals (0.81 to 0.82), suggesting that recent data and long-term deviations moderately influence rainfall prediction. Five-period trends (0.80 to 0.81) and three-period rolling means (0.79 to 0.80) further contribute to predictive power, though to a lesser extent. The seven-period trend, with the lowest correlation values (0.72 to 0.74), is the least effective predictor. Thus, integrating short-term trends, rolling means, and deviations at multiple timescales into predictive models can significantly enhance rainfall forecast accuracy.

C. LINEAR INTERPOLATION

Linear interpolation is used to estimate missing values based on the values before and after the missing entries [29].

The interpolation method can handle various other interpolation methods (e.g., polynomial, spline), but linear is a common choice for simplicity and efficiency. Linear interpolation fills in missing values by using the equation 1 for each missing value:

$$y_i = y_{i-1} + (y_{i+1} - y_{i-1}) \times \frac{x_i - x_{i-1}}{x_{i+1} - x_{i-1}} \quad (1)$$

where:

y_i is the missing value to be interpolated.

y_{i-1} and y_{i+1} are the known values immediately before and after the missing value.

x_i is the position of the missing value.

x_{i-1} and x_{i+1} are the positions of the known values before and after the missing value.

IV. HYPERPARAMETER SELECTION

Hyperparameter selection is crucial in machine learning because the performance of models often heavily depends on the choice of hyperparameters.

TABLE 6- SELECTED HYPER PARAMETERS

Model Name	Hyper Parameter Name	Value
CNN	Epochs	200
	Kernel Size	2
	Drop Out	0.1
	Filters (Conv1D)	32, 64, 64
	Pool Size (MaxPooling1D)	2
	Dense Layer Size	64
	Loss Function	Mean Squared Error (MSE)
	Activation Function	Adam
LSTM	Epochs	200
	Number of LSTM Units	64
	Number of Dense Units	128
	Activation Function	Tanh
XGB	Eta	0.17
	Max Depth	15
	Min Child Weight	15.81
	Subsample	0.67
	ColsampleBytree	0.75
	Gamma	0.36
	Lambda	1.87
	Alpha	0.91
CNN-XGB	The same parameters selected for CNN and XGB	

Traditional methods like grid search or random search can be inefficient and time-consuming, especially when dealing with large or complex models. Optuna addresses this by automating the search and employing more sophisticated strategies [30]. Optuna is an open-source, automatic hyperparameter optimization framework designed to efficiently and flexibly tune hyperparameters in machine learning models. It provides a user-friendly interface to optimize complex objective functions, which makes it popular among data scientists and machine learning practitioners. Table 6 presents the chosen hyperparameters for this study.

V. METHODOLOGY

The research methodology, depicted in Fig. 4, describes the steps for training and evaluating a hybrid CNN-XGBoost model. The process starts with data preprocessing, which involves normalizing, splitting the dataset, and statistical feature engineering. Following this, the CNN extracts important features, which are then combined with the original features to create Combined Data. This Combined Data is used to train the XGBoost model. The model's accuracy and performance are then evaluated by testing it on a separate test dataset, with performance metrics including MAE, RMSE, and R2.

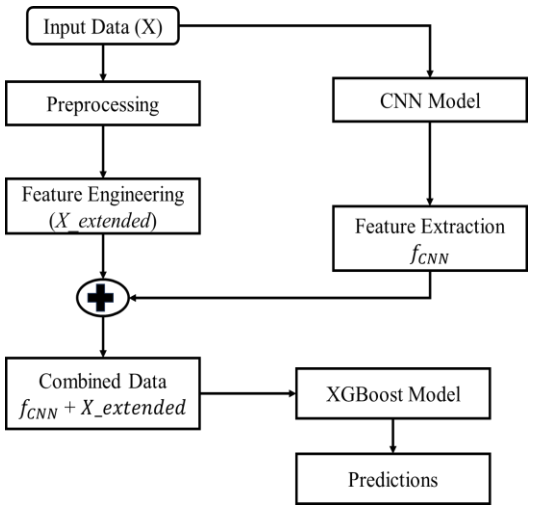

FIGURE 4. **Overall flowchart of proposed methodology**

A. PROPOSED CNN-XGBOOST MODEL

The proposed hybrid model combines features extracted from a CNN with an XGBoost model to predict rainfall. The key steps and the necessary equations are discussed below.

Input Data: The input shape is (T,1) where T is the number of time steps.

1. Convolutional Neural Network (CNN)

Because Convolutional Neural Networks (CNNs) can automatically extract and learn essential elements from huge and complicated weather datasets, they have become an effective tool for rainfall prediction. CNNs can capture the geographical and temporal correlations essential for precise rainfall predictions by processing time series data from several observation locations.

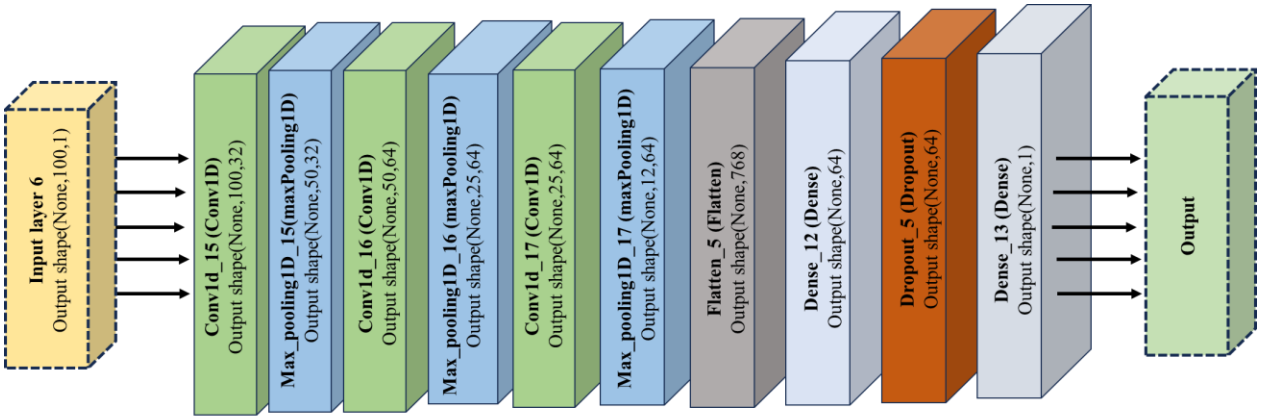

FIGURE 5. **Architecture of CNN model**

These networks are particularly good at seeing patterns in a wide range of meteorological variables, including wind speed, humidity, and air temperature—all of which significantly impact rainfall. CNNs can identify minute differences and patterns in the data that conventional statistical techniques can overlook because of the convolutional layers inside of them [31].

Furthermore, by using methods like data augmentation and interpolation, CNNs can deal with noise and missing data, which are frequent problems in weather datasets. They are robust for real-world applications because of their capacity to handle segmented and normalized data, which guarantees that they can generalize well to novel and unseen scenarios. CNNs are versatile in some meteorological scenarios because they can be trained to anticipate rainfall across a variety of time horizons, from short-term forecasts to longer-range projections. CNNs have demonstrated great potential in enhancing the precision of rainfall forecasts in areas with extremely dynamic weather patterns [32], [33].

Additionally, CNNs can reach high accuracy rates and low error metrics by utilizing optimization algorithms like Adam, which makes them dependable for usage in operational settings. Their use in meteorology improves prediction skills and helps weather-sensitive industries prepare for disasters and manage resources more effectively. All things considered, CNNs are a breakthrough in rainfall prediction, offering a sophisticated method for deciphering and predicting intricate meteorological occurrences [34].

The detailed architecture of the CNN model within the hybrid model is illustrated in Fig. 5, and the step-by-step process for this model, along with the corresponding mathematical equations, is thoroughly discussed below:

i. Convolution operation with a filter K: The input data X is convolved with a filter KK to detect patterns. The convolution operation can be described as:

$$(X*K)(i,j)=\sum_{u=0}^{m-1}\sum_{v=0}^{n-1} X\,(\,i+u, j+v\,)\cdot K\,(\,u,v\,) \quad (2)$$

Here, X is the input data, K is the filter, (i,j) are the coordinates of the output feature map, and (u,v) are the coordinates of the filter.

ii. Activation function: The **ReLU** (Rectified Linear Unit) activation function introduces non-linearity into the model, which helps in learning complex patterns. It is defined as:

$$ReLU\,(X) = max\,(0, X) \quad (3)$$

iii. Pooling layer: The **Max-Pooling** operation reduces the spatial dimensions of the feature maps while retaining the most important features. It is defined as:

$$P\,(i, j) = max\,\{X\,(i+a,j+b\,)\,/0 \le a < p,\, 0 \le b < q\,\} \quad (4)$$

where, $P(i,j)$ is the pooled feature map, and (a,b) are the dimensions of the pooling window.

iv. Flatten and fully connected layers: The pooled feature maps are flattened into a single vector and passed through one or more fully connected layers. This step combines the extracted features to form a high-level representation:

$$y = \sigma\,(W_f + b) \quad (5)$$

Here, W is the weight matrix, f is the flattened feature vector, b is the bias term, and σ is an activation function such as sigmoid or softmax.

v. Output layer of CNN: The output layer predicts the rainfall amount based on the features extracted by the CNN.

$$\hat{R} = w^T y + b \quad (6)$$

Here, w is the weight vector, and b is the bias term.

2. Feature extraction from CNN: The pooled feature maps P_i are flattened and passed through fully connected layers to form the final feature vector f_{CNN}.

$$f_{CNN} = CNN_{features}(X) = \sigma\,(W \cdot Flatten(P) + b) \quad (7)$$

3. Combine CNN features with original input: Concatenate the original features X_{scaled} with the CNN-extracted features f_{CNN} to form the hybrid feature vector f_{hybrid}. The hybrid feature vector f_{hybrid} has dimensions $N\times(F+d)$, where N is the number of samples, F is the number of original features, and d is the number of CNN-extracted features.

$$f_{hybrid} = [\,X_{scaled}, f_{CNN}] \quad (8)$$

$$f_{hybrid} \in R^{N\times(F+d)} \quad (9)$$

4. XGBoost model

XGBoost is a powerful machine-learning algorithm that has been effectively used for rainfall prediction due to its ability to handle structured data with high accuracy. It is an ensemble learning technique based on decision trees that creates a strong predictive model by combining the predictions of several weak learners. Because it can simulate intricate interactions between different meteorological variables, such as temperature, humidity, and wind speed, which are essential for precise forecasts, XGBoost is especially well-suited for rainfall prediction. Its capacity to deal with missing data—a frequent problem in meteorological datasets—by utilizing built-in techniques to estimate missing values and preserve forecast performance is one of its main advantages [35], [36].

The algorithm's gradient boosting framework allows it to minimize prediction errors through iterative refinement, making it highly efficient in capturing non-linear patterns in the data. XGBoost also provides flexibility in feature selection, automatically identifying the most relevant features from a large set of weather variables, which enhances the overall model accuracy. Its regularization techniques prevent overfitting, ensuring that the model generalizes well to new, unseen data, which is crucial for reliable rainfall prediction [37].

Also, XGBoost can simply be adjusted to maximize performance, is extremely scalable, and can handle enormous datasets that are common in meteorology. Because of its efficiency and speed, it can be used for real-time prediction applications where precise forecasts must be made quickly. Overall, XGBoost is a useful tool for rainfall prediction because of its resilience and versatility; it provides accurate and dependable forecasts that can help with resource planning and crisis management [38]. Overall architecture of XGBoost model is demonstrated in Fig. 6.

i. Training XGBoostmodel: Create a DMatrix for training using the hybrid feature vector and the true rainfall amounts y. Train the XGBoost model on this DMatrix to learn the relationship between the features and the target variable.

$$dtrain = xgb.DMatrix\,(f_{hybrid}, label = y) \quad (10)$$

Where $y \in R^N$ is the true rainfall amount

ii. Final prediction: Use the trained XGBoost model to predict rainfall amounts for the test data. Create a DMatrix for the test data in the same way as for the training data.

$$\hat{R}xgb = xgb.predict\,(dtest) \quad (11)$$

where, X_{test_hybrid} is the matrix of test data features, and d_{test} is the corresponding DMatrix. The prediction for each test sample is obtained by summing the predictions from all trees in the XGBoost ensemble

$$\hat{R}_i(xgb) = \sum_{t=1}^{T} f_t(x_{i,test\_hybrid}) \quad (12)$$

$\hat{R}_i(xgb)$ is the predicted rainfall amount for the *i-th* test sample. $f_t(x_{i,test\_hybrid})$ is the prediction from the *t-th* tree in the ensemble for the *i-th* test sample $x_{i,test\_hybrid}$. Thus, for all test samples, the predicted vector is $\hat{R}_i(xgb)$ where $\hat{R}_i\,(xgb) \in R^{N_{test}}$. Mathematical equation of $\hat{R}_i\,(xgb)$ is shown in equation 13.

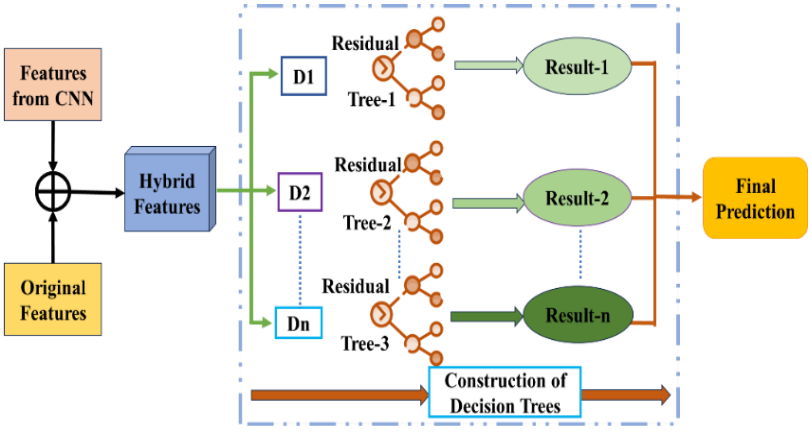

FIGURE 6. **Architecture of XGBoost model**

$$\hat{R}_i\,(xgb) = \begin{bmatrix} \hat{R}_1(xgb) \\ \hat{R}_2(xgb) \\ . \\ . \\ \hat{R}_{N_{test}}\,(xgb) \end{bmatrix} = \begin{bmatrix} \sum_{t=1}^{T} f_t(x_{1,test\_hybrid}) \\ \sum_{t=1}^{T} f_t(x_{2,test\_hybrid}) \\ . \\ . \\ \sum_{t=1}^{T} f_t(x_{N_{test},test\_hybrid}) \end{bmatrix} \quad (13)$$

VI. RESULT

A. EVALUATION CRITERIA

In the context of machine learning, evaluation criteria are numerical measurements that are used to evaluate a prediction model's efficacy and performance [39]. These standards shed light on the model's effectiveness in carrying out its intended function. Different regression metrics like mean squared error (MSE) or R-squared are examples of common evaluation criteria. The selection of criteria is contingent upon the nature of the problem (classification versus regression) and particular objectives (e.g., fraud detection: minimizing false positives). For a thorough assessment of the model's performance, it is imperative to choose relevant criteria that match the demands of the problem and, where needed, to combine several metrics.
Mathematical equation for MAE, RMSE, and R2 are shown below in Equation 14, 15 and, 16:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |\, y_i - \hat{y}_i \,| \quad (14)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (|\, y_i - \hat{y}_i \,|)^2} \quad (15)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (|y_i - \hat{y}_i|)^2}{\sum_{i=1}^{n} (|y_i - \hat{y}_i|)^2} \quad (16)$$

Where n is the number of samples or observations, y_i represents the actual values, $\hat{y}_i$ represents the predicted values.

B. FOR SATELLITE DATA

Table 7 provides a comprehensive comparison of several predictive models—CNN, LSTM, XGB, CNN-XGB, and Transformer-XGB—across multiple districts (Rajshahi, Bogura, Chapai, Joypurhat, Naogoan, Natore, Pabna, and Sirajganj). The models are evaluated based on three performance metrics: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R2).

TABLE 7- EVALUATION METRICS BY DIFFERENT MODELS ON THE SATELLITE DATASET

District	Model	RMSE	MAE	R2
Rajshahi	CNN	1.07	0.63	0.98
	LSTM	0.88	0.39	0.99
	XGB	0.98	0.45	0.98
	Transformer-XGB	0.96	0.46	0.98
	CNN-XGB	**0.65**	**0.28**	**0.99**
Bogura	CNN	1.16	0.61	0.99
	LSTM	1.84	0.87	0.97
	XGB	1.73	0.72	0.97
	Transformer-XGB	1.74	0.74	0.97
	CNN-XGB	**0.97**	**0.42**	**0.99**
Chapaina wabganj	CNN	0.65	0.36	0.99
	LSTM	0.95	0.48	0.98
	XGB	1.14	0.41	0.97
	Transformer-XGB	1.15	0.4	0.97
	CNN-XGB	**0.88**	**0.25**	**0.98**
Joypurhat	CNN	1.33	0.69	0.97
	LSTM	1.10	0.49	0.98
	XGB	1.13	0.53	0.98
	Transformer-XGB	1.18	0.55	0.98
	CNN-XGB	**0.82**	**0.33**	**0.99**
Naogoan	CNN	1.14	0.67	0.98
	LSTM	1.07	0.45	0.98
	XGB	1.15	0.51	0.98
	Transformer-XGB	1.17	0.54	0.98
	CNN-XGB	**0.85**	**0.35**	**0.99**
Natore	CNN	1.80	0.99	0.95
	LSTM	0.87	0.41	0.99
	XGB	1.03	0.5	0.98
	Transformer-XGB	1.15	0.54	0.98
	CNN-XGB	**0.62**	**0.3**	**0.99**
Pabna	CNN	1.35	0.83	0.98
	LSTM	1.56	0.79	0.98
	XGB	1.67	0.76	0.97
	Transformer-XGB	1.61	0.63	0.98
	CNN-XGB	**1.21**	**0.49**	**0.99**
Sirajganj	CNN	2.32	1.09	0.97
	LSTM	2.21	0.89	0.97
	XGB	1.99	0.85	0.97
	Transformer-XGB	1.97	0.89	0.98
	CNN-XGB	1.32	0.51	0.99

In Rajshahi, the CNN-XGB model outperforms the other models with an RMSE of 0.65, an MAE of 0.28, and an R2 of 0.99, indicating high accuracy and low error. The LSTM model also performs well, with an RMSE of 0.88, an MAE of 0.39, and an R2 of 0.99, suggesting it is a strong contender. Traditional models like CNN and XGB show higher errors compared to the hybrid models.

The CNN-XGB model again shows superior performance in Bogura, achieving an RMSE of 0.97, an MAE of 0.42, and an R2 of 0.99. The standalone CNN model performs well with an R2 of 0.99, but its RMSE and MAE are higher at 1.16 and 0.61, respectively. LSTM and Transformer-XGB models have higher errors, with LSTM having the highest RMSE of 1.84.

In Chapainawabganj, the CNN-XGB model continues to excel with an RMSE of 0.88, an MAE of 0.25, and an R2 of 0.98. The CNN model performs admirably as well, with an RMSE of 0.65 and an R2 of 0.99, though its MAE is slightly higher at 0.36. Transformer-XGB and XGB models show comparatively higher errors.

For Joypurhat, the CNN-XGB model is the best performer with an RMSE of 0.82, an MAE of 0.33, and an R2 of 0.99. The LSTM and XGB models also perform well, with both achieving an R2 of 0.98, but the CNN-XGB model's lower error rates make it the top choice.

In Naogoan, the CNN-XGB model again leads with an RMSE of 0.85, an MAE of 0.35, and an R2 of 0.99. Both the CNN and LSTM models show good performance, with R2 values of 0.98, but higher RMSE and MAE values compared to the hybrid model.

The CNN-XGB model outperforms all other models in Natore, achieving an RMSE of 0.62, an MAE of 0.30, and an R2 of 0.99. The LSTM model also performs well with an RMSE of 0.87 and an R2 of 0.99. Other models, including CNN and XGB, show significantly higher errors.

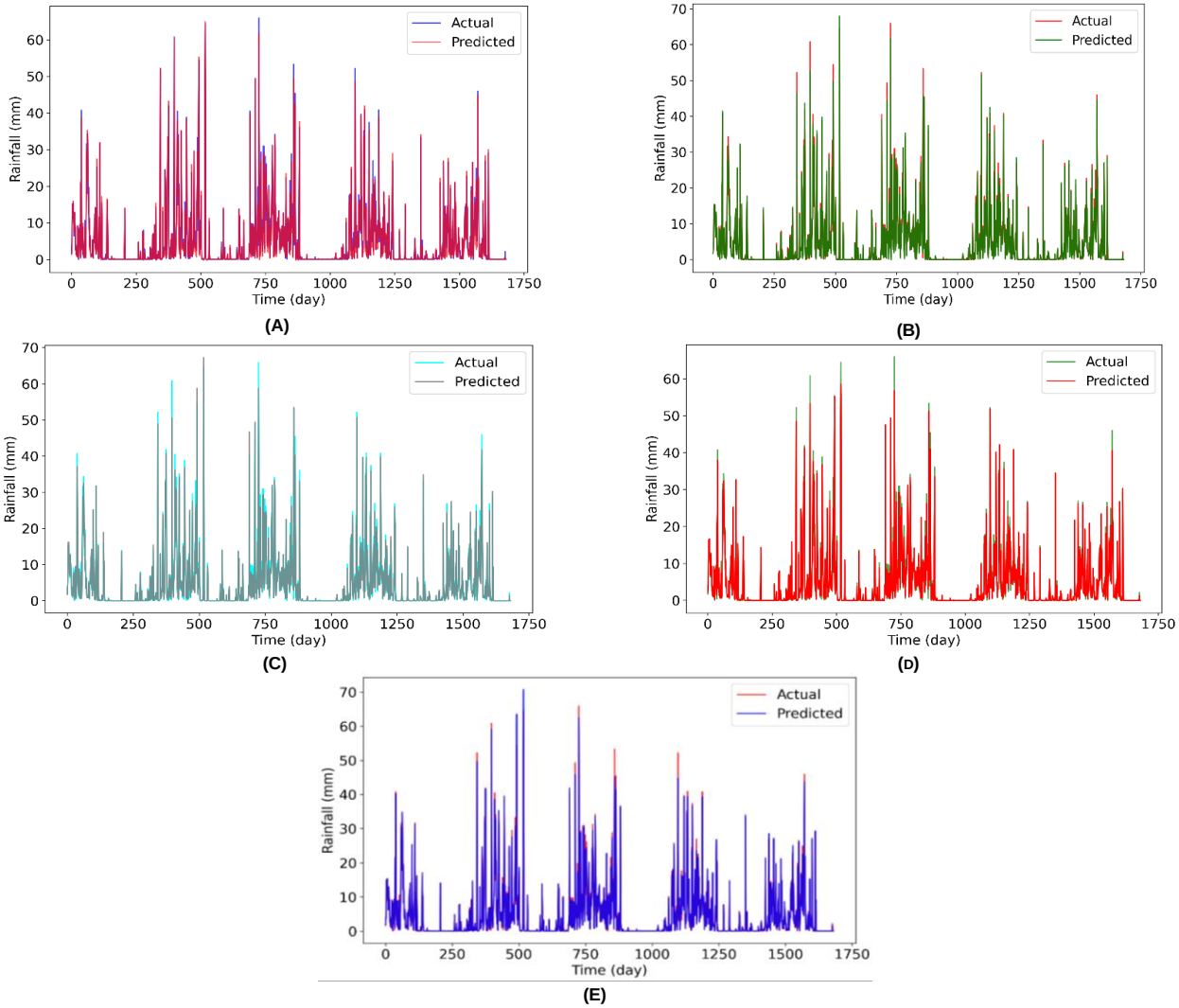

FIGURE 7. Actual vs predicted rainfall by (A) LSTM, (B) CNN, (C) XGBoost, (D) XGBoost-Transformer, and (E) CNN-XGBoost (proposed) model for full test data (Rajshahi district).

In Pabna, the CNN-XGB model shows strong performance with an RMSE of 1.21, an MAE of 0.49, and an R2 of 0.99. The CNN and LSTM models also perform well with R2 values of 0.98 but have higher RMSE and MAE values compared to the hybrid model. The CNN-XGB model achieves the best performance in Sirajganj with an RMSE of 1.32, an MAE of 0.51, and an R2 of 0.99. The LSTM and Transformer-XGB models show relatively high errors, with RMSE values of 2.21 and 1.97, respectively.

Across all districts, the CNN-XGB hybrid model consistently demonstrates superior performance with the lowest RMSE and MAE values and the highest R2 values, indicating it is the most accurate and reliable model for predicting the given dataset. The standalone CNN and LSTM models also perform well but generally show higher error rates compared to the CNN-XGB hybrid. This suggests that hybrid models, combining the strengths of different algorithms, provide significant improvements in predictive accuracy and error reduction.

Fig.7 compares the performance of five models— LSTM, CNN, XGBoost, XGBoost-Transformer, and CNN-XGBoost—on predicting signal values for the Rajshahi district's full test data. Each graph illustrates the actual versus the predicted rainfall values. Every testing model capture the general trend but struggles with high fluctuations, leading to some discrepancies during signal peaks. The extent of these struggles varies between models, and based on Table 7, it is evident that the proposed model captures the signal peaks more accurately than the others.

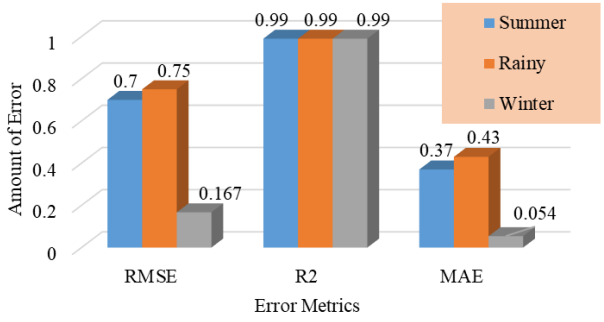

FIGURE 8. Season-wise comparison of different error metrics for the proposed model.

Fig. 8 demonstrates the performance metrics of a predictive model across three different seasons—Summer, Rainy, and Winter. Across all seasons, the model demonstrates a very high R^2 value of 0.99, indicating that the model explains 99% of the variance in the data, signifying robust predictive ability. Seasons do affect the average size of mistakes, as measured by the RMSE and MAE.

The model performs best in Winter, with the lowest RMSE (0.167) and MAE (0.054) which suggests the proposed model's predictions are most accurate during this season. In contrast, the model's performance is slightly less accurate in the Summer and Rainy seasons, with RMSE values of 0.7 and 0.75, and MAE values of 0.37 and 0.43, respectively. This variance indicates that although the model fits the data well across all seasons, it performs best in the winter season in terms of accuracy and precision in forecasting the exact values.

Fig. 9 illustrates a comparative analysis between actual recorded rainfall and predicted rainfall over a period spanning from July 1, 2021, to October 1, 2021. The graph includes three lines: the actual rainfall data (blue), predicted rainfall data (red), and the prediction error (gray), which represents the difference between the actual and predicted values.

The close alignment between the actual and predicted rainfall lines indicates the model's high accuracy in forecasting rainfall during the analyzed period. However, the model struggles to accurately capture abrupt changes, leading to some prediction errors.

The ground dataset was specifically utilized to assess the robustness of the proposed model. Consequently, we focused solely on presenting the error metrics derived from this dataset, as it served as a critical measure of the model's performance under challenging conditions. Given that the primary purpose of using this dataset was to evaluate the model's stability and reliability, we did not include figures illustrating the overall test data or seasonal performance. This decision was made to maintain a clear emphasis on the robustness evaluation rather than on broader performance trends.

C. FOR GROUND DATA

Table 8 presents a comparative analysis of the performance metrics—RMSE, MAE, and R^2—across five models (CNN, LSTM, XGB, Transformer-XGB, and CNN-XGB) applied to four districts: Rajshahi, Bogura, Pabna, and Dinajpur. In Rajshahi, the CNN-XGB model outperformed other models with an RMSE of 16.28, an MAE of 7.85, and an R^2 of 0.98, indicating superior predictive accuracy. The CNN model also performed well with an RMSE of 17.00, an MAE of 7.85, and an R^2 of 0.97, while the LSTM model had a slightly higher RMSE of 17.82 and an MAE of 9.63, but maintained an R^2 of 0.97. The Transformer-XGB model showed relatively lower performance with an RMSE of 21.11, an MAE of 12.84, and an R^2 of 0.96. In Bogura, the CNN-XGB model once again led with an RMSE of 22.60, an MAE of 8.42, and an R^2 of 0.98, followed closely by the CNN model, which recorded an RMSE of 22.64, an MAE of 8.70, and an R^2 of 0.97. The XGB model had a similar R^2 of 0.98 but showed a higher RMSE of 22.92 and a significantly higher MAE of 13.68. The Transformer-XGB model, however, had the weakest performance in this district, with an RMSE of 28.77, an MAE of 14.68, and an R^2 of 0.96. In Pabna, the CNN-XGB model again excelled with an RMSE of 19.55, an MAE of 10.09, and an R^2 of 0.98. The CNN model followed with an RMSE of 21.37, an MAE of 11.35, and an R^2 of 0.97. The LSTM model, however, exhibited lower performance, with an RMSE of 29.46, an MAE of 13.99, and an R^2 of 0.94.

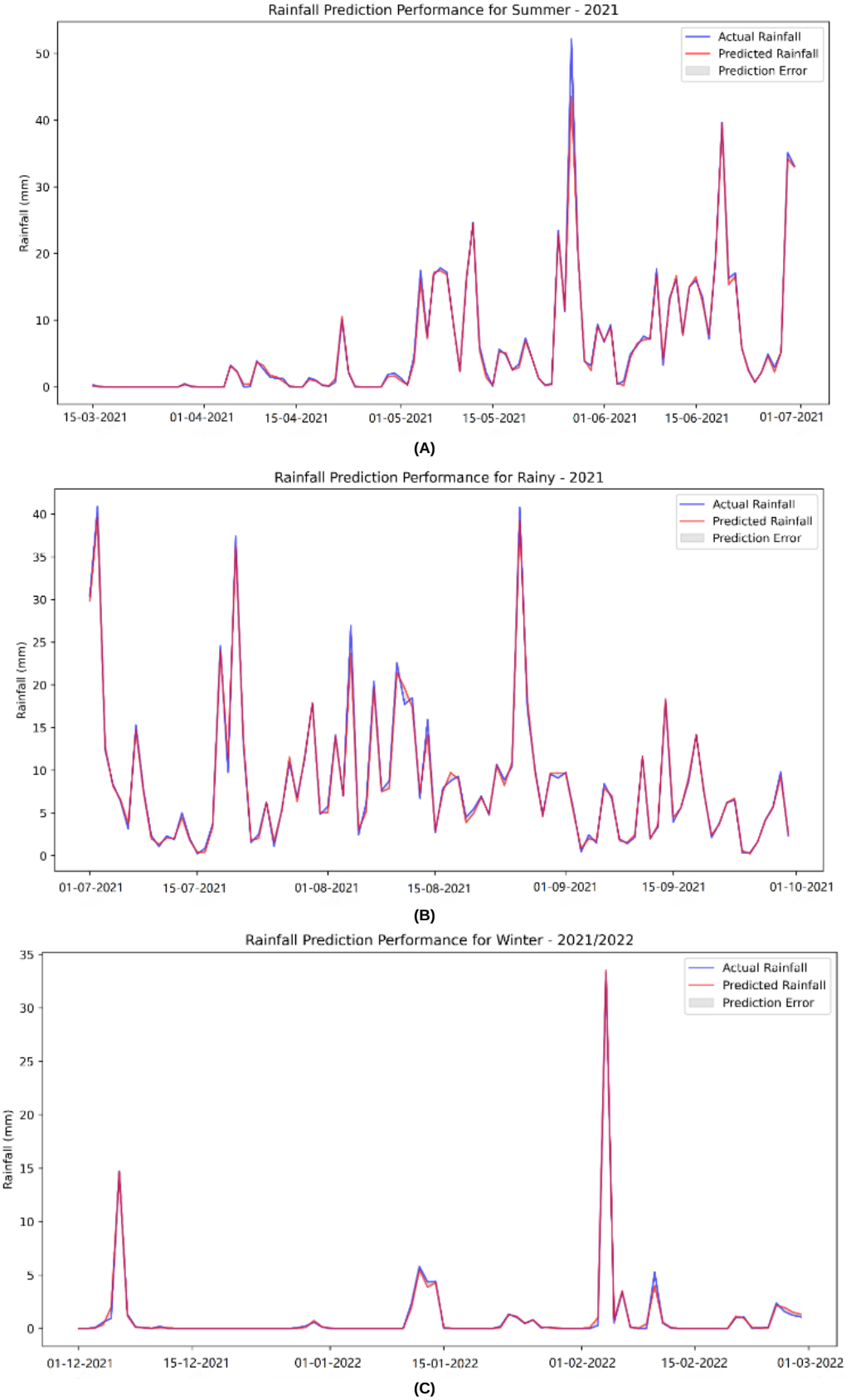

FIGURE 9. Actual vs predicted rainfall for (A) Summer season 2021, (B) Rainy season 2021, and (C) Winter season 2021-22

Finally, in Dinajpur, while the CNN-XGB model maintained a strong R^2 of 0.97, its RMSE and MAE increased to 26.03 and 10.40, respectively. The CNN model in Dinajpur recorded an RMSE of 24.68, an MAE of 8.21, and an R^2 of 0.97, showcasing competitive performance among the models.

TABLE 8- EVALUATION METRICS BY DIFFERENT MODELS ON THE GROUND DATASET

District	Model	RMSE	MAE	R2
Rajshahi	CNN	17.00	7.85	0.97
	LSTM	17.82	9.63	0.97
	XGB	17.38	10.67	0.98
	Transformer-XGB	21.11	12.84	0.96
	CNN-XGB	16.28	7.85	0.98
Bogura	CNN	22.64	8.7	0.97
	LSTM	24.09	9.35	0.97
	XGB	22.92	13.68	0.98
	Transformer-XGB	28.77	14.68	0.96
	CNN-XGB	22.60	8.42	0.98
Pabna	CNN	21.37	11.35	0.97
	LSTM	29.46	13.99	0.94
	XGB	23.22	14.04	0.97
	Transformer-XGB	23.09	13.29	0.96
	CNN-XGB	19.55	10.09	0.98
Dinajpur	CNN	24.68	8.21	0.97
	LSTM	30.61	12.46	0.96
	XGB	31.02	14.57	0.96
	Transformer-XGB	31.51	15.64	0.96
	CNN-XGB	26.03	10.40	0.97

D. COMPARISION OF RESULTS

The difference in error metrics between Table 7 and Table 8 can be attributed to the complexity and richness of the datasets used in each case. In Table 7, the satellite dataset comprises 16 variables, providing a comprehensive set of features for the models to learn from. This multi-variable dataset allows the models to capture complex relationships and patterns, leading to lower error metrics such as RMSE, MAE, and higher R^2 values across various districts. The diversity of input features enhances the models' ability to predict rainfall with greater accuracy, as they can leverage correlations among multiple factors.

On the other hand, Table 8 presents the evaluation metrics for the ground dataset, which is limited to a single variable: rainfall (measured in mm/month). The simplicity of this dataset, with only one input feature, significantly constrains the models' ability to capture the full complexity of rainfall prediction. As a result, the models exhibit higher error metrics, indicating reduced predictive accuracy. The absence of additional variables to cross-reference and correlate means the models must rely solely on the historical rainfall data, which limits their robustness and leads to a greater degree of error.

From Table 9, the substantial differences in RMSE (15.63) and MAE (7.57) for Rajshahi between the satellite and ground datasets clearly demonstrate the impact of dataset richness on model performance. The satellite dataset, with its 16 variables, allows the model to capture more variance in the rainfall data, leading to a slightly higher R2. This indicates that the model trained on the satellite dataset has a marginally better fit to the data compared to the model trained on the ground dataset. The small difference in R2 (0.01) is consistent with the differences observed in RMSE and MAE, reinforcing the idea that a more comprehensive dataset enhances model performance by allowing it to better capture the underlying patterns in the data.

TABLE 9- COMPARISON OF ERROR METRICS BETWEEN SATELLITE AND GROUND DATASETS FOR RAJSHAHI DISTRICT

Dataset	RMSE	MAE	R2
Satellite	0.65	0.28	0.99
Ground	16.28	7.85	0.98
Difference	15.63	7.57	0.01

VII. EXPLAINABLE AI

SHAP values provide a unified measure of feature importance for any machine learning model. They are based on Shapley values from cooperative game theory, ensuring that the contributions of each feature are fairly distributed according to their marginal contributions across all possible feature subsets [35]. The equation for the SHAP value of a feature i in an instance x is shown in equation 17.

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\cdot(|N|-|S|-1)!}{|N|!}\left[f_x(S\cup\{i\}) - f_x(S)\right] \quad (17)$$

where:

ϕ_i SHAP value for feature i.

N: Set of all features.

S: Subset of features not including i.

|S|: Number of features in subset S.

|N|: Total number of features.

$f_x(S)$: Model's prediction for instance x using features in subset S.

$f_x(S \cup \{i\})$: Model's prediction for instance x using features in subset S plus feature i.

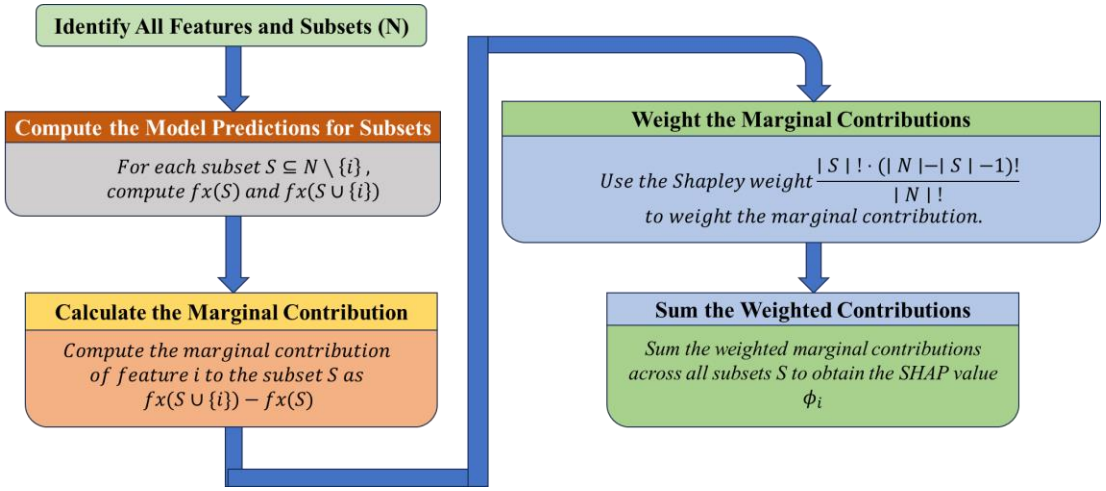

FIGURE 10. **Step-by-Step Process of Calculation of SHAP Values**

A. SHAP FORCE PLOT FOR THE PROPOSED MODEL

This section aims to interpret the decision-making process of our proposed Hybrid CNN-XGBoost model for Rajshahi Division rainfall prediction using explainable AI with Shapley Additive exPlanations (SHAP). The term "explainable AI" (XAI) describes strategies and tactics that enable people to understand the results and workings of machine learning (ML) models.

XAI seeks to provide understandable, comprehensible explanations for decision-making processes, in contrast to conventional "black box" ML models that offer minimal insight into their decision-making processes. The explanation module in Explainable AI (XAI) could identify which input features influence the model's output the most. Such explanations are only useful when the model's output corresponds to the actual, accurate outcome [40]. SHAP (Shapley Additive exPlanations) is a machine learning approach for explaining model predictions. It is based on game theory concepts, especially Shapley values, which determine the impact of each feature on a model's outcome. SHAP works by breaking down the model's predictions into the total influence of each feature. This method assigns a numerical value to each feature, reflecting its impact on the model's ultimate prediction. These values not only assist in determining which features are most influential, but they also provide insights into how the model makes decisions. This interpretability is critical for maintaining transparency, confidence, and the practical deployment of machine learning models across multiple domains [41].

We employed SHAP in our proposed hybrid CNN XGBoost model to better understand its behavior and boost its dependability.The CNN processes 18 raw attributes and extracts 64 new ones. These extracted features are combined with the raw data to get 82 hybrid features fed into the XGBoost model. SHAP values aid in interpreting both models, resulting in two types of plots: mean absolute attribute values for global feature significance and bee swarm plots to demonstrate the influence of top features on output. SHAP values for the CNN reveal which raw features contribute the most to the extracted features, whereas SHAP values for XGBoost show which hybrid features contribute the most to the overall prediction.

Fig.11(a) below shows the SHAP Mean Absolute Attribute (MAA) values for the top twenty impacting features in the CNN model used for rainfall prediction. The x-axis shows the mean absolute SHAP value, which indicates the average influence of each feature on the model's predictions, while the y-axis displays the features. The feature PRECTOTCORR trend2 is the most impactful, with a mean SHAP value of +1.24, suggesting the largest average influence on the CNN model's predictions. Following closely are PRECTOTCORR residual 14 (+1.09), PRECTOTCORR residual (+0.71), and PRECTOTCORR ewm mean2 (+0.70), all of which provide considerable contributions to the model's output. Other important characteristics are PRECTOTCORR trend3 (+0.49), PRECTOTCORR rolling mean2 (+0.43), and PRECTOTCORR trend7 (+0.33). In contrast, meteorological parameters such as QV2M (+0.12), T2M (+0.010), WD10M (+0.08), WD2M (+0.08), RH2M (+0.04), and TS (+0.02) have significantly lesser impacts. This graphic successfully demonstrates which attributes the CNN model depends on the most when making predictions, with higher SHAP values suggesting more relevance.

The provided bee swarm representation of Fig. 11(b) depicts the SHAP values for the CNN model's feature extraction procedure using 18 raw features, demonstrating how these factors influence rainfall prediction. Each dot represents a SHAP value for a feature in a single prediction, and the color indicates the feature value (blue for low values, red for high values). The x-axis displays the SHAP value, which represents the influence of each attribute. The plot shows how low and high values affect rainfall estimates for each characteristic. The features with the greatest SHAP values

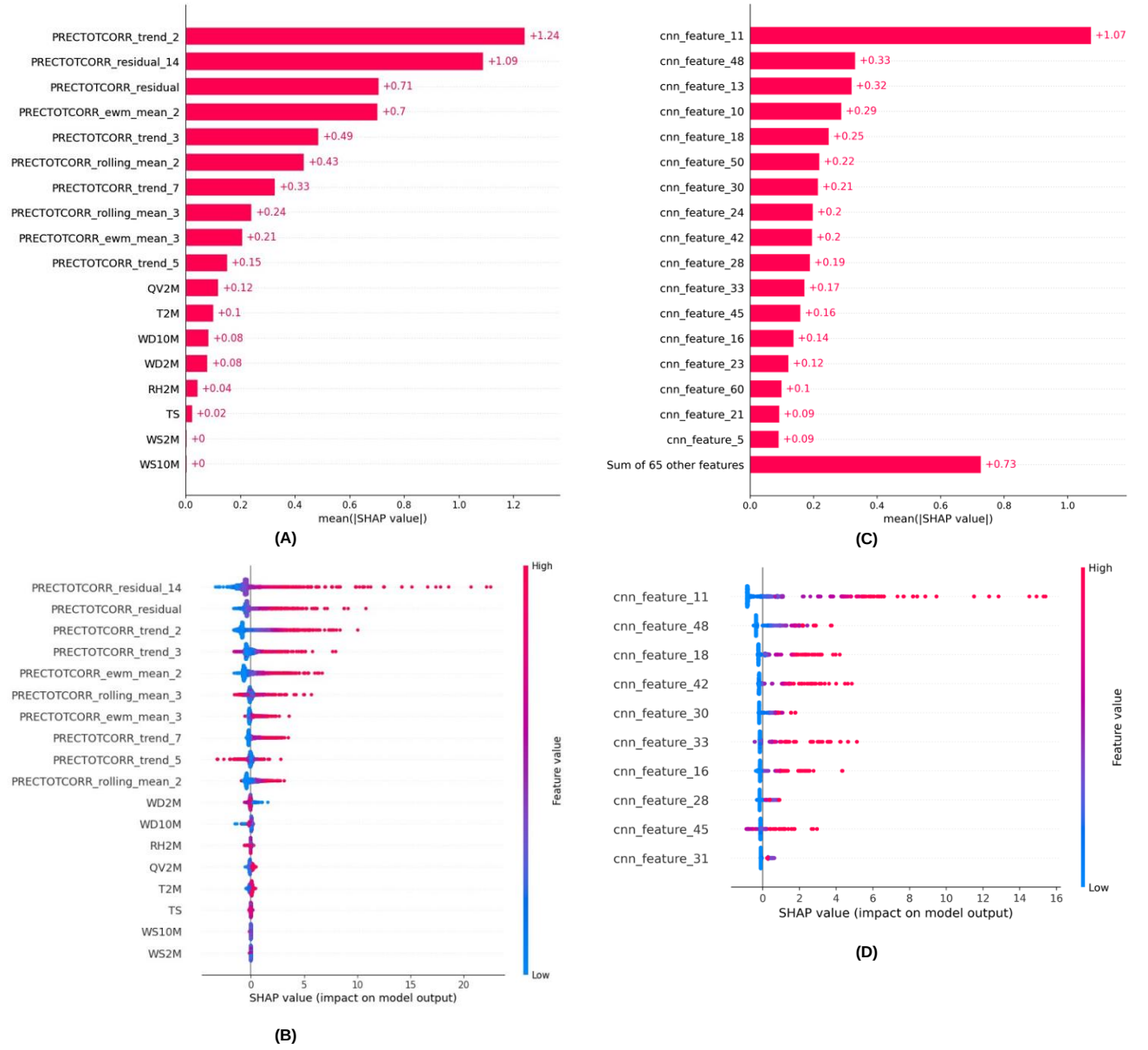

FIGURE 11. (A) Bar Plot of CNN model, (B) Bee Swarm Plot of CNN model, (C) Bar Plot of nested XGBoost Model, and (D) Bee Swarm Plot of nested XGBoost Model

are PRECTOTCORR residual 14, PRECTOTCORR residual, and PRECTOTCORR trend 2, showing a considerable effect. High values (red) of PRECTOTCORR residual14, PRECTOTCORR residual, PRECTOTCORR trend 2, and other attributes improve rainfall forecast, whereas low levels (blue) have the opposite effect. This pattern is consistent with other characteristics, including PRECTOTCORR trend 3, PRECTOTCORR ewm mean 2, and PRECTOTCORR rolling mean 3. Similarly, high values of meteorological parameters such as WD2M, WD10M, RH2M, QV2M, T2M, TS, WS10M, and WS2M improve rainfall prediction, but low values have the opposite effect. This graphic successfully emphasizes the relevance and unpredictability of individual raw features in CNN's feature extraction process, illustrating how changing feature values have a positive or negative impact on rainfall forecasts.

Fig. 11(c) depicts the SHAP Mean Absolute Attribute (MAA) values for the XGBoost model, which employs 81 hybrid features created by blending 18 raw features and 64 CNN-extracted features. The x-axis shows the mean absolute SHAP value, which indicates the average influence of each hybrid feature on the model's rainfall forecast.

At the same time, the y-axis displays the characteristics in descending order of relevance. cnn feature 11 has the highest mean SHAP value of 1.07, suggesting that it has the greatest average impact on the final predictions. Following it are cnn feature 48 and cnn feature 13, which have mean SHAP values of 0.33 and 0.32, respectively. Other prominent features include cnn feature 10, cnn feature 18, and cnn

feature 50, contributing considerably to the model's output. A cumulative impact from the remaining 65 features is shown, with a collective mean SHAP value of 0.73. The following chart clearly shows which hybrid traits the XGBoost model depends on the most for accurate rainfall forecasts, with larger SHAP values suggesting greater relevance and impact. The chart clearly shows that among the hybrid features affecting the final output, the most influential features are primarily those extracted by CNN. A feature like cnn_feature_11 stands out with the highest mean SHAP value of 1.07, indicating that it has the most significant average influence on the XGBoost model's predictions. Furthermore, the cumulative contribution of the remaining 65 features, which include all raw data, is represented by a collective mean SHAP value of 0.73, suggesting that these raw features have the least influence on the final rainfall forecast.

Fig. 11(d) depicts a bee swarm plot that summarizes the contribution of 10 parameters from the top twenty in the final prediction model. In our SHAP value study for rainfall prediction, the top 10 CNN-extracted characteristics had various levels of effect. cnn_feature_11 has the most influence, with high values contributing considerably favorably and low values reducing its impact. Similarly, cnn_feature_48 has a significant influence, with greater values resulting in positive contributions and lower ones producing negative consequences. The features cnn feature 18, cnn feature 42, cnn feature 30, cnn feature 33, cnn feature 16, cnn feature 28, cnn feature 45, and cnn feature 31 exhibit similar behavior, with high values favorably impacting predictions and low values adversely influencing them. Overall, greater values of these CNN characteristics improve the model's predictions, but lower values degrade accuracy. This summary illustrates the main effects of the top ten characteristics out of the top twenty in our rainfall prediction model.

VIII. DISCUSSION

This study assessed the performance of various machine learning models for predicting rainfall in several districts of Bangladesh, using both satellite and ground datasets. Overall, the hybrid CNN-XGB model proved to be the most successful. In most areas, including Rajshahi, Bogura, Chapai, and Sirajganj, CNN-XGB scored the highest R^2 values of 0.99 for the satellite dataset, demonstrating its superior accuracy and capacity to simulate intricate rainfall patterns. With the ground dataset, especially in Bogura, Pabna, and Dinajpur, the model showed similarly impressive results, obtaining R^2 of 0.98 and RMSE values as low as 19.00. These results imply that CNN-XGB has a very high degree of reliability when predicting rainfall in various situations.

The results show how well the CNN-XGB model represents intricate spatial relationships and nonlinear interactions in rainfall data. The hybrid model, which combines Convolutional Neural Networks (CNN) and Extreme Gradient Boosting (XGB), performs consistently well in forecasting rainfall across a range of datasets and geographical areas. Its resilience is further demonstrated by the low RMSE and MAE values, which make it a useful tool for real-world scenarios. This suggests that CNN-XGB can accurately anticipate rainfall in various conditions, which is important for strategic planning.

For Bangladesh, where agriculture, water management, and disaster preparedness rely on accurate forecasts, rainfall prediction is essential. The CNN-XGB model's better performance has important ramifications for improving these industries. The model can help minimize agricultural losses, enhance water management, and prepare for weather-related calamities by offering accurate rainfall projections. According to this study, hybrid models have the potential to significantly increase forecast accuracy not just in Bangladesh but also in other areas with comparable environmental difficulties.

The study's geographic scope is limited to just a few areas in Bangladesh, which might not adequately represent the range of climate conditions seen around the globe. Although efficient, the hybrid CNN-XGB model requires a lot of computing power, which may limit its implementation in real-time scenarios lacking sophisticated resources. Also, the long-term forecasting potential was not explored because the main focus was on short-term projections.

Future research should consider expanding the scope of this study by testing the models across various climatic regions, such as semi-arid and Mediterranean zones, to assess their adaptability and robustness. It is necessary to create and assess long-term rainfall prediction models to assess their suitability for use in strategic planning. Investigating different hybrid strategies, such as ensemble methods or stacking, might provide extra gains in prediction accuracy. To facilitate real-time deployment in operational forecasting, efforts should be made to maximize the computing efficiency of these models. Other exogenous variables that could be included to improve model predictions and provide a more thorough knowledge of rainfall dynamics include soil moisture and socioeconomic characteristics. The CNN-XGB model's proven efficacy highlights its potential to revolutionize weather forecasting and offer crucial assistance for agricultural, water management, and disaster preparedness decision-making.

A. PRACTICAL IMPLEMENTATIONS

Superior performance of CNN-XGB model in rainfall prediction has significant implications for urban planning, particularly in regions prone to riverbank erosion and flooding. Urban planners can utilize these accurate rainfall forecasts to design infrastructure that mitigates flood risks and enhances the resilience of urban areas. By predicting heavy rainfall events with high accuracy, city planners can implement preemptive measures such as improving drainage

systems, constructing flood barriers, and designing green spaces that absorb excess water. This is especially crucial for Bangladesh, where rapid urbanization along riverbanks increases the vulnerability of urban areas to erosion and flooding.

The detailed and accurate forecasts provided by the CNN-XGB model can also guide the placement and design of critical infrastructure, such as roads, bridges, and housing, to minimize damage during extreme weather events. Additionally, urban planners can use these predictions to optimize the allocation of resources for emergency response and disaster preparedness, ensuring that areas most at risk of flooding receive timely attention and support. By integrating the model's forecasts into urban planning processes, cities can develop more resilient infrastructures that are better equipped to handle the challenges posed by climate change and extreme weather events.

Furthermore, the ability of the CNN-XGB model to simulate intricate rainfall patterns and nonlinear interactions in the data provides a deeper understanding of the spatial and temporal distribution of rainfall. This information is invaluable for developing zoning regulations that prevent construction in high-risk areas, such as floodplains and regions vulnerable to riverbank erosion. Urban planners can also use these insights to create more effective land-use plans that balance development with the need to preserve natural floodplains, thereby reducing the overall risk of flood damage.

IX. CONCLUSION

This study developed and compared various rainfall prediction models using five distinct machine learning and deep learning algorithms, with two different types of data inputs. Rainfall time series data from satellite observations for eight districts in Bangladesh—Rajshahi, Bogura, Chapainawabganj, Joypurhat, Naogoan, Natore, Pabna, Sirajganj—and from four ground stations in the northern region, including Rangpur and Sylhet, were utilized for training and testing. The algorithms employed were CNN, LSTM, XGB, Transformer-XGB, and CNN-XGB. Model parameters were optimized using the Optuna algorithm. The performance of each algorithm was evaluated using three evaluation metrics: coefficient of determination (R^2), mean absolute error (MAE), and root mean square error (RMSE). The main findings from the study are as follows:

The CNN-XGB model, especially in comparison to other models, delivered accurate and consistent precipitation predictions for both data sources without any significant drop in performance. This highlights the robustness of the proposed model.

The satellite dataset's 16 variables significantly improved model performance, reducing RMSE by 15.63 and MAE by 7.57, and slightly increasing R2 by 0.01. This emphasizes the importance of a comprehensive dataset for accurate and reliable predictions.

Utilizing explainable algorithms such as SHAP is crucial for building trust and providing transparency in the use of ML models, thereby supporting the adoption of data-driven approaches in rainfall early warning systems. By employing the SHAP explainer, this study identified the significance of climatic variables, feature engineering, and features generated by the CNN model as predictors for the XGB model in rainfall prediction. This was confirmed across various rainfall conditions and events, affirming their inclusion in the proposed model.

This research focused on precipitation modeling in two divisions in northern Bangladesh. Additionally, no clear relationship was established between the data from adjacent districts or regions. Future studies should extend to other areas with tropical monsoon climates as well as regions with different climatic features, such as Mediterranean and semi-arid climates. These studies should also explore the relationships between adjacent regions to improve the accuracy of rainfall predictions.

ACKNOWLEDGMENT This research work is supported by 30-196-2-058 from Qatar Research, Development and Innovation (QRDI). The authors also would like to thank Qatar National library (QNL) for their funding and assistance in the open access publica tion. The authors would like to acknowledge the use of AI systems, specifically ChatGPT, Grammarly, and Quillbot, for language improvement, editing, and grammar enhancement.

REFERENCES

[1] R. He, L. Zhang, and A. W. Z. Chew, "Data-driven multi-step prediction and analysis of monthly rainfall using explainable deep learning," *Expert Syst Appl*, vol. 235, Jan. 2024, doi: 10.1016/j.eswa.2023.121160.

[2] K. Bansal, A. K. Tripathi, A. C. Pandey, and V. Sharma, "RfGanNet: An efficient rainfall prediction method for India and its clustered regions using RfGan and deep convolutional neural networks," *Expert Syst Appl*, vol. 235, Jan. 2024, doi: 10.1016/j.eswa.2023.121191.

[3] F. Di Nunno, F. Granata, Q. B. Pham, and G. de Marinis, "Precipitation Forecasting in Northern Bangladesh Using a Hybrid Machine Learning Model," *Sustainability (Switzerland)*, vol. 14, no. 5, Mar. 2022, doi: 10.3390/su14052663.

[4] K. Johny, M. L. Pai, and A. S., "A multivariate EMD-LSTM model aided with Time Dependent Intrinsic Cross-Correlation for monthly rainfall prediction," *Appl Soft Comput*, vol. 123, Jul. 2022, doi: 10.1016/j.asoc.2022.108941.

[5] A. Y. Barrera-Animas, L. O. Oyedele, M. Bilal, T. D. Akinosho, J. M. D. Delgado, and L. A. Akanbi, "Rainfall prediction: A comparative analysis of modern machine learning algorithms for time-series forecasting," *Machine*

Learning with Applications, vol. 7, p. 100204, Mar. 2022, doi: 10.1016/j.mlwa.2021.100204.

[6] S. Miau and W. H. Hung, "River flooding forecasting and anomaly detection based on deep learning," *IEEE Access*, vol. 8, pp. 198384–198402, 2020, doi: 10.1109/ACCESS.2020.3034875.

[7] X. H. Le, D. H. Nguyen, S. Jung, M. Yeon, and G. Lee, "Comparison of Deep Learning Techniques for River Streamflow Forecasting," *IEEE Access*, vol. 9, pp. 71805–71820, 2021, doi: 10.1109/ACCESS.2021.3077703.

[8] R. N. Iyengar and S. T. G. R. Kanth, "Intrinsic mode functions and a strategy for forecasting Indian monsoon rainfall." [Online]. Available: www.tropmet.res.in

[9] Y. Zhao *et al.*, "AI-based rainfall prediction model for debris flows," *Eng Geol*, vol. 296, Jan. 2022, doi: 10.1016/j.enggeo.2021.106456.

[10] A. Mahabub, A.-Z. Sultan, and B. Habib, "An Overview of Weather Forecasting for Bangladesh Using Machine Learning Techniques."

[11] T. woong Kim, H. Ahn, G. Chung, and C. Yoo, "Stochastic multi-site generation of daily rainfall occurrence in south Florida," *Stochastic Environmental Research and Risk Assessment*, vol. 22, no. 6, pp. 705–717, 2008, doi: 10.1007/s00477-007-0180-8.

[12] Y. Zaman, "Machine Learning Model on Rainfall-A Predicted Approach for Bangladesh," 2018.

[13] Q. Zou, Y. Liu, and X. Linge, "A survey on rainfall forecasting using artificial neural network'," 2019.

[14] I. Hossain, H. M. Rasel, M. A. Imteaz, and F. Mekanik, "Long-term seasonal rainfall forecasting using linear and non-linear modelling approaches: a case study for Western Australia," *Meteorology and Atmospheric Physics*, vol. 132, no. 1, pp. 131–141, Feb. 2020, doi: 10.1007/s00703-019-00679-4.

[15] L. Ni *et al.*, "Streamflow and rainfall forecasting by two long short-term memory-based models," *J Hydrol (Amst)*, vol. 583, Apr. 2020, doi: 10.1016/j.jhydrol.2019.124296.

[16] S. Poornima and M. Pushpalatha, "Prediction of rainfall using intensified LSTM based recurrent Neural Network with Weighted Linear Units," *Atmosphere (Basel)*, vol. 10, no. 11, Nov. 2019, doi: 10.3390/atmos10110668.

[17] P. Singh and B. Borah, "Indian summer monsoon rainfall prediction using artificial neural network," *Stochastic Environmental Research and Risk Assessment*, vol. 27, no. 7, pp. 1585–1599, Oct. 2013, doi: 10.1007/s00477-013-0695-0.

[18] Y. Xiang, L. Gou, L. He, S. Xia, and W. Wang, "A SVR–ANN combined model based on ensemble EMD for rainfall prediction," *Applied Soft Computing Journal*, vol. 73, pp. 874–883, Dec. 2018, doi: 10.1016/j.asoc.2018.09.018.

[19] S. Adarsh and M. J. Reddy, "Multiscale characterization and prediction of monsoon rainfall in India using Hilbert–Huang transform and time-dependent intrinsic correlation analysis," *Meteorology and Atmospheric Physics*, vol. 130, no. 6, pp. 667–688, Dec. 2018, doi: 10.1007/s00703-017-0545-6.

[20] K. Johny, M. L. Pai, and S. Adarsh, "Adaptive EEMD-ANN hybrid model for Indian summer monsoon rainfall forecasting," *Theor Appl Climatol*, vol. 141, no. 1–2, pp. 1–17, Jul. 2020, doi: 10.1007/s00704-020-03177-5.

[21] M. I. Khan and R. Maity, "Hybrid Deep Learning Approach for Multi-Step-Ahead Daily Rainfall Prediction Using GCM Simulations," *IEEE Access*, vol. 8, pp. 52774–52784, 2020, doi: 10.1109/ACCESS.2020.2980977.

[22] A. G. Salman, Y. Heryadi, E. Abdurahman, and W. Suparta, "Weather forecasting using merged Long Short-Term Memory Model (LSTM) and Autoregressive Integrated Moving Average (ARIMA) Model," *Journal of Computer Science*, vol. 14, no. 7, pp. 930–938, 2018, doi: 10.3844/jcssp.2018.930.938.

[23] G. J. Sawale and S. R. Gupta, "Use of Artificial Neural Network in Data Mining For Weather Forecasting," *International Journal Of Computer Science And Applications*, vol. 6, no. 2, 2013, [Online]. Available: www.researchpublications.org

[24] C. L. Wu, K. W. Chau, and C. Fan, "Prediction of rainfall time series using modular artificial neural networks coupled with data-preprocessing techniques," *J Hydrol (Amst)*, vol. 389, no. 1–2, pp. 146–167, Jul. 2010, doi: 10.1016/j.jhydrol.2010.05.040.

[25] F. R. Adaryani, S. Jamshid Mousavi, and F. Jafari, "Short-term rainfall forecasting using machine learning-based approaches of PSO-SVR, LSTM and CNN," *J Hydrol (Amst)*, vol. 614, Nov. 2022, doi: 10.1016/j.jhydrol.2022.128463.

[26] S. Cramer, M. Kampouridis, A. A. Freitas, and A. K. Alexandridis, "An extensive evaluation of seven machine learning methods for rainfall prediction in weather derivatives," *Expert Syst Appl*, vol. 85, pp. 169–181, Nov. 2017, doi: 10.1016/j.eswa.2017.05.029.

[27] "POWER | DAV." Accessed: Aug. 13, 2024. [Online]. Available: https://power.larc.nasa.gov/data-access-viewer/

[28] E. A. Madrid and N. Antonio, "Short-Term Electricity Load Forecasting with Machine Learning," *Information 2021, Vol. 12, Page 50*, vol. 12, no. 2, p. 50, Jan. 2021, doi: 10.3390/INFO12020050.

[29] T. Niedzielski and M. Halicki, "Improving Linear Interpolation of Missing Hydrological Data by Applying Integrated Autoregressive Models," *Water Resources Management*, vol. 37, no. 14, pp. 5707–5724, Nov. 2023, doi: 10.1007/S11269-023-03625-7/FIGURES/9.

[30] P. S. Pravin, J. Z. M. Tan, K. S. Yap, and Z. Wu, "Hyperparameter optimization strategies for machine learning-based stochastic energy efficient scheduling in cyber-physical production systems," *Digital Chemical Engineering*, vol. 4, p. 100047, Sep. 2022, doi: 10.1016/J.DCHE.2022.100047.

[31] M. Qiu *et al.*, "A short-term rainfall prediction model using multi-task convolutional neural networks," in *Proceedings - IEEE International Conference on Data Mining, ICDM*, Institute of Electrical and Electronics

Engineers Inc., Dec. 2017, pp. 395–404. doi: 10.1109/ICDM.2017.49.
[32] Y. R. Sari, E. C. Djamal, and F. Nugraha, "Daily Rainfall Prediction Using One Dimensional Convolutional Neural Networks," in *2020 3rd International Conference on Computer and Informatics Engineering, IC2IE 2020*, Institute of Electrical and Electronics Engineers Inc., Sep. 2020, pp. 90–95. doi: 10.1109/IC2IE50715.2020.9274572.
[33] S. P. Van, H. M. Le, D. V. Thanh, T. D. Dang, H. H. Loc, and D. T. Anh, "Deep learning convolutional neural network in rainfall-runoff modelling," *Journal of Hydroinformatics*, vol. 22, no. 3, pp. 541–561, May 2020, doi: 10.2166/hydro.2020.095.
[34] A. Haidar and B. Verma, "Monthly Rainfall Forecasting Using One-Dimensional Deep Convolutional Neural Network," *IEEE Access*, vol. 6, pp. 69053–69063, 2018, doi: 10.1109/ACCESS.2018.2880044.
[35] A. Ibrahem Ahmed Osman, A. Najah Ahmed, M. F. Chow, Y. Feng Huang, and A. El-Shafie, "Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia," *Ain Shams Engineering Journal*, vol. 12, no. 2, pp. 1545–1556, Jun. 2021, doi: 10.1016/j.asej.2020.11.011.
[36] M. T. Anwar, E. Winarno, W. Hadikurniawati, and M. Novita, "Rainfall prediction using Extreme Gradient Boosting," in *Journal of Physics: Conference Series*, IOP Publishing Ltd, Apr. 2021. doi: 10.1088/1742-6596/1869/1/012078.
[37] R. G. Sanches, P. Augusto, T. Rios, and R. G. Sanches, "Using Xgboost Models For Daily Rainfall Prediction." [Online]. Available: https://ssrn.com/abstract=4778138
[38] M. Ma *et al.*, "XGBoost-based method for flash flood risk assessment," *J Hydrol (Amst)*, vol. 598, Jul. 2021, doi: 10.1016/j.jhydrol.2021.126382.
[39] K. SerefogluCabuk*et al.*, "Chasing the objective upper eyelid symmetry formula; R2, RMSE, POC, MAE, and MSE," *Int Ophthalmol*, vol. 44, no. 1, pp. 1–9, Dec. 2024, doi: 10.1007/S10792-024-03157-Y/METRICS.
[40] R. He, L. Zhang, and A. W. Z. Chew, "Data-driven multi-step prediction and analysis of monthly rainfall using explainable deep learning," *Expert Syst Appl*, vol. 235, Jan. 2024, doi: 10.1016/j.eswa.2023.121160.
[41] G. Van Den Broeck, A. Lykov, M. Schleich, and D. Suciu, "On the Tractability of SHAP Explanations," 2022.

MD. SAFAYET ISLAM received his B.Sc. degree in electrical and computer engineering from Rajshahi University of Engineering & Technology (RUET), Rajshahi, Bangladesh, in 2022. He worked as a Quantitative Analyst at Anchorblock Technologies, a fintech company, where he conducted research on data analysis, data mining, and high-frequency trading. His work involved developing and applying algorithms for algorithmic trading, conducting time series analysis and decomposition, and researching the implementation of machine learning techniques on real-time financial market data. His research interests include machine learning and its applications in financial markets, power sector optimization, environmental analysis, and image classification.

MD SHAFIUZZAMAN received the B.Sc. degree in electrical and computer engineering from the Rajshahi University of Engineering and Technology, Rajshahi, Bangladesh, in 2022. He is currently serving as a Lecturer in the Department of Computer Science and Engineering at North Bengal International University. He won the poster presentation at the International Conference on Advances and Challenges through Translational Research in Biological Sciences, where he presented a poster titled "IoT-Based Smart Soil Fertilizer Monitoring, Suggestion, and Crop Disease Detection System Using Deep Learning." His research interests include machine learning applications in time series forecasting, energy management, environmental and agricultural systems, focusing on electrical load forecasting, rainfall prediction with explainable AI, and image classification.

GOLAM MAHMUD received the B.Sc. degree in electrical and computer engineering from Rahshahi University of Engineering & Technology (RUET), Rajshahi, Bangladesh, on October 8, 2023. Previously, he worked as an artificial intelligence engineer at Deep Mind Labs Ltd. to build a recommendation system. Currently, he serves as a computer science and engineering faculty member at Dhaka International University (DIU), Dhaka, Bangladesh. His research interests include computer vision, time series forecasting, and natural language processing.

NABILA NOWSHIN is a final year student of Electrical and Computer Engineering at Rajshahi University of Engineering and Technology (RUET), Rajshahi, Bangladesh. She is currently deepening her knowledge in machine learning, with a keen interest in its applications in data analysis, time series forecsting, image processing, and artificial intelligence.

PARISA REZA currently pursuing B.Sc. degree in electrical and computer engineering in Rajshahi University of Engineering and Technology (RUET), Bangladesh. Recently she has been learning machine learning technology. Her interest in field of research is time series forecasting, fraud detection focusing on cybercrime detection integrated with machine learning.

JAHID HASAN is a young Urban Planner by profession. He graduated from Urban & Regional Planning from Rajshahi University of Engineering & Technology (RUET). Currently, he is serving as a faculty member in the same department. He is also an active Associate Member of the Bangladesh Institute of Planners (BIP). His research interests include the integration of public transit and active transportation, transportation policy and planning, and the application of GIS in transportation planning.

MD. NAHIDUZZAMAN is currently pursuing PhD in Computer Science at RMIT University, Australia. He received his M.Sc. and B.Sc. degrees in Computer Science and Engineering from Rajshahi University of Engineering and Technology (RUET), Rajshahi, Bangladesh, in 2018 and 2023 respectively. He is currently an Assistant Professor in the Department of Electrical and Computer Engineering at RUET. Additionally, he was a Research Assistant with the Qatar University Machine Learning Research Group. He has several peer-reviewed journal publications. His research interests include machine learning and its applications in disease detection, the power sector, and agriculture.

AMITH KHANDAKAR (Senior Member, IEEE) received the B.Sc. degree in electronics and telecommunication engineering from North South University, Bangladesh, the master's degree in computing (networking concentration) from Qatar University, in 2014, and the Ph.D. degree in biomedical engineering from Universiti Kebangsaan Malaysia (UKM), Malaysia, in 2023. He has currently around 100 journal publications, ten book chapters, and three registered patents under his name. His research interests include sensors and instrumentation, electronics, engineering education, biomedical engineering, and machine learning applications. He is a certified Project Management Professional and the Cisco Certified Network Administrator. He graduated as the Valedicto rian (President Gold Medal Recipient) of North South University.

MOHAMED ARSELENE AYARI received the Ph.D. degree from the Department of Civil and Environmental Engineering, University of Col orado at Boulder, with a focus on building systems and energy management. He is currently an Asso ciate Professor with the Department of Civil and Architectural Engineering, Qatar University. He is also affiliated with the Technology Innovation and Engineering Education (TIEE) Program. He is act ing as a PI in three research grants: one is an NPRP-cluster project promoting sustainable development of K12 STEM education in Qatar in a digital age and two are QU internal grants: Managed Aquifer Recharge (MAR) in Qatar: A Multiscale Investigation of Clogging of Qatari Groundwater Aquifers in HIG Cycle 05, and a porous-scale investigation of the dynamics of contact angles for geological carbon sequestration in HIG Cycle 06. Among his latest research publications are ''Perspectives on Food Waste Management: Prevention and Social Innovations'' published in the journal Sustainable Production and Consumption; ''Automatic and Reli able Leaf Disease Detection Using Deep Learning Techniques'' published in the journal Agri Engineering; ''Fouling Mitigation Strategies for Different Foulants in Membrane Distillation'' published in the journal Chemical Engineering and Processing-Process Intensification; and ''Estimating the Relative Crystallinity of Biodegradable Polylactic Acid and Polyglycolide Polymer Composites by Machine Learning Methodologies'' published in the Polymers Journal. His multidisciplinary areas of research interests include bioengineering, contaminants distribution, indoor air quality, and computational thermofluidic dynamics (CFD).