cognizancejournal.com
Shahran Rahman Alve *et al*, Cognizance Journal of Multidisciplinary Studies, Vol.4, Issue.8, August 2024, pg. 274-289
(An Open Accessible, Multidisciplinary, Fully Refereed and Peer Reviewed Journal)
ISSN: 0976-7797
Impact Factor: 4.843
Index Copernicus Value (ICV) = 77.57

Harnessing Historical Weather Data with Machine Learning for Rainfall Forecasting in Bangladesh: A Time Series Analysis

Shahran Rahman Alve[1*]; Al Jubayer Pial[1]; Muhammad Zawad Mahmud[1]; Samiha Islam[1]

[1]Department of Electrical and Computer Engineering, North South University, Bashundhara, Dhaka-1229, Bangladesh
[*]Corresponding Author: Shahran Rahman Alve, Email: shahran.alve@northsouth.edu
shahran.alve@northsouth.edu; jubayer.pial@northsouth.edu;
zawad.mahmud1@northsouth.edu; samiha.islam2@northsouth.edu

DOI: 10.47760/cognizance.2024.v04i08.017

ABSTRACT:
This study aims to explore the integration of historical weather data with machine learning methods for better rainfall forecasting in Bangladesh. It has the potential to identify rainfall early by observing their prediction on preventing natural disasters. Weather forecasting, a process that is paramount to protecting human lives and property as well as general welfare, at times faces challenges due mainly in part to manual computations. These calculations are likely error-ridden because of the vastness and complexity of data needing to be analyzed. This research combines weather forecasting with machine learning techniques to improve the accuracy and precision of predictions. The main goal is to develop an accurate prediction system of rainfall forecasting in Bangladesh for the future. The model training data set is from historical meteorologic records of Bangladesh during the period 1901–2015. After neatly hyper-tuning the Random Forest model, we achieve some impressive RMSE metrics. The results are compared with a baseline model, Randomized Search CV, using different evaluation metrics such as Mean Absolute Error (MAE), Mean Squared Logarithmic Error (MSLE), and Root-mean-squared-log-error (RMSLE), etc., and the observations of significantly better performance is observable. Machine learning in weather forecasting has become an essential tool that can save lives and billions of dollars in property. An extensive analysis of machine learning techniques was performed in the domain of rainfall prediction. Among all the tried models, the Random Forest model gave perfect predictions, as noted by its Mean Squared Error (MSE) of 12245.52 and absolute error value of 64.25. Furthermore, there was excellent precision (0.95) and recall (0.92), respectively, for the classification of wet days, corresponding to a total accuracy score of 91%. Based on the above result, we can conclude that the Random Forest model is both robust and promising for rainfall prediction problems. Various models such as Linear Regression, Support Vector Regressor (SVR), Decision Tree, K-Nearest Neighbors (KNN), AdaBoost, XG Boost, Ridge, Linear SVR and MLP Regressor showed variations in prediction accuracies and errors. Nevertheless, the Random Forest came out as a better choice in this case, showing its superiority. The plan is to expand the time period of the dataset along with more advance machine learning models to predict weather forecast in future.

Keywords: Weather; Rainfall; Machine Learning; Hydrological Forecasting; Time Series Analysis.

cognizancejournal.com
Shahran Rahman Alve *et al*, Cognizance Journal of Multidisciplinary Studies, Vol.4, Issue.8, August 2024, pg. 274-289
(An Open Accessible, Multidisciplinary, Fully Refereed and Peer Reviewed Journal)
ISSN: 0976-7797
Impact Factor: 4.843
Index Copernicus Value (ICV) = 77.57

Introduction

Background and Motivation

The occurrence of heavy rainfall events can have numerous disastrous consequences. For instance, it can make floods likely to occur, thereby promoting the rapid spread of waterborne and vector-borne diseases. At the same time, floods can cause severe damage to the fields and crops of people who do not have insurance on major personal property items (homes), as well as causing a life threat. Moreover, intense precipitation events often significantly affect road traffic systems and daily life. In the middle of 2019, the Northern part of Bangladesh faced severe fallout due to a sudden flood precipitated by heavy rainfall [1]. As a result, the frequency of cholera, diarrhea, dysentery and skin diseases, as well as other waterborne illnesses, has increased. Around 500,000 hectares of potential farmland became unsuitable for cultivation, in addition to around 600,000 residential units and nearly 7,000 km (4,300 miles) of transport infrastructure being distorted or demolished. The flood killed 75, according to the same source [2]. In addition, when unfavorable weather conditions persist over a long period, and there is too little rain, the effects of drought occur, which can then lead to solid decreases in food production [3]. The hydro-meteorological dynamics in South Asia — especially pertaining to Bangladesh, present a very somber picturization of climate-induced extremities. In 2020, when the theme was "restoring wetlands", the monsoons left about a quarter of the nation underwater and severely damaged agricultural lands as well as infrastructures [2], with floods that affected around 4.7 million people during this time, aggravated by heavy rains which were present for far too long. A similar pattern was observed in 2021, with more than 190000 hectares of croplands affected during the year causing severe threats to food security and livelihood primarily because agro-chemical inputs failed due to herbicide-resistant weed species spreading across farming systems [3]. Specially, the 2022 floods were especially devastating in the northeastern Sylhet division, where historic rainfall displaced hundreds of thousands and killed many [4]. A slight improvement was seen in 2023 but the confluence of floods every year made it an overwhelming task and reflected a need for advanced prediction mechanisms [5].

The devastation that the floods of 2024 had left behind was a grim reminder of this story. In 2024, Bangladesh experiences one of the worst floods in its history. It was just another example of the impact that extreme weather events, made more severe by climate change, wreak upon millions and countless other domestic fowl — dislocating communities with traffic virtually impossible, leading to widespread destruction or loss of infrastructure such as roads and agriculture. This resulted in floods of a ferocity never seen before by human or natural landscapes, inundating millions and revealing how fragile existing infrastructural paradigms are. This year's flood reaffirmed that acute meteorological anomalies can still hash out tremendous damage —extending, for instance, to leaving more than an area of communal farming land occupied by starved communities— but also exposed the ingrained systemic enemy in favor of such regular disasters. A need for improved predictive models became clear with the flood of 2024 which killed many and caused extensive damages to property, livelihoods. The consequences of high temperatures also have a wide range whereby this affects the rate in which illness and death occur due to health concerns. Excessive cold or heat can cause body aches and other problems like heat stroke and diarrhea. This is especially true for senior citizens and children who spend much of the day in an outdoor environment [6]. We saw a glimpse of this during the severe heatwave in Southern Bangladesh in mid-last year. Similarly, at the end of 2019, cold waves led to respiratory infections and gastrointestinal syndromes, causing death among a total of about fifty people [7]. This significantly strained healthcare facilities, which suddenly found themselves treating dehydration, pneumonia, and influenza cases. An effect on enzyme activity in plants and soil, on the moisture content of the soil or its respiration rate may result from this temperature deviation [6]. Even a few degrees difference can cause significant crop loss, and that, in turn, could hurt the agricultural industry. Because the agriculture industry plays such a prominent role in the national economy, negative ramifications could result from extremes of too much precipitation and abnormal temperatures. And then it has the possibility of a rise in death tolls. This was reflected in the agricultural losses with a value of BDT 11.52 billion due to the flood that occurred in Bangladesh in 2019, making on-time and certain weather forecasting an ultimatum for potential mitigation against such huge economic damages [1]. Preventative measures alike to flood protection strategies and water resource management are necessary so as to reduce these impacts.

cognizancejournal.com
Shahran Rahman Alve *et al*, Cognizance Journal of Multidisciplinary Studies, Vol.4, Issue.8, August 2024, pg. 274-289
(An Open Accessible, Multidisciplinary, Fully Refereed and Peer Reviewed Journal)
ISSN: 0976-7797
Impact Factor: 4.843
Index Copernicus Value (ICV) = 77.57

Purpose and Goal of the Project

Predicting the temperature and rainfall patterns with precision is quite challenging because these phenomena are inherently non-linear. Prior prediction systems using theoretical models limited model quality and computational resources. Conversely, using data-driven approaches to reproduce predictions has been shown to achieve orders of magnitude improvements in accuracy. These models are built using mathematical, statistical methodologies and machine learning. Statistical techniques are considered fundamental in weather prediction research. Statistical and machine learning methods are the most common tools for building data-driven models. In weather prediction research, statistical methodologies have played a significant role in the past. The Holt-Winters additive method was used to predict rainfall in [4] and the ARIMA model used for forecasting of monsoon precipitation especially at Allahabad city only [5]. On the other hand, more sophisticated models such as statistical methods provide poor performance when are applied to non-linear pattern-based phenomenon due to their limitations on this aspect. Non-linear approaches performed better in weather prediction studies, and this has made the use of predictive machine learning techniques a more attractive choice for research work. As already mentioned, both Artificial Neural Networks (ANNs) and Gene Expression Programming (GEP) have been used in the past for rainfall estimation, with ANNs more popular than GEP. Rainfall forecasting has been a research area in particular areas, and machine learning techniques such as single-layer feedforward Networks (SLFN) have been applied. However, as time evolves with events in a recursive manner, it requires the consideration of past estimation states and, therefore, is an issue for machine learning methods commonly seen today. One of the most popular Deep Learning (DL) techniques used is Recurrent Neural Networks (RNNs), as they combine temporal data. RNN (Recurrent Neural Networks), in addition to having a good capability for capturing temporal dependencies, run into vanishing gradient problems when forced to capture long-term dependencies. Due to this constraint, Long Short-Term Memory (LSTM) networks were developed to solve the vanishing gradients problem. The Long Short-Term Memory (LSTM) model has proven to be effective in many fields, including financial market forecasting [8], natural language processing, wind turbine damage detection [9] and air quality prediction. Deep learning techniques have become widespread due to their low computational complexity and automatic feature extraction capabilities, making them useful in various fields.

The primary goal of this research is to develop a rainfall forecasting system with higher accuracies by evaluating the performance of different machine learning models on a real-time basis of historical weather data for Bangladesh over a considerably long timeframe (1901–2015). The aim is to evaluate the best model of rainfall forecasting and its predictive accuracy. To achieve this, several models are assessed, including RandomForestRegressor, DecisionTreeRegressor, LinearRegressor, SVR, KNeighborsRegressor, MLPRegressor, AdaBoostRegressor, and GradientBoostingRegressor. There are some metrics to evaluate the performance of models, such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared, to assess the predictability and accuracy of each model. In comparing these various metrics, the objective is to use this information to ascertain which model would likely represent better for rainfall prediction and would improve general forecasting quality.

Research Literature Review

Existing Research and Limitations

Developments in rainfall predictions can be transformed into more robust models when we combine them with Machine Learning (ML) and time-series analysis techniques, especially as climate change appears to bring about increased unpredictability. Barrera-Animas et al. (2021), a critical study in this space by [11] which employed a comparative analysis of multiple ML models, namely LSTM and Stacked-LSTM; Bidirectional-LSTM Networks; XGBoost, as well as ensemble approaches using data from the past two decades for major UK cities. The results follow the experiment, which shows that graph attention-based architectures outperformed LSTM-based ones. The bidirectional-lstm-network was surprising in demonstrating compatibility with much fewer layers compared to stacked-lstms. This research not only underscores the capability of simple ML models to reduce computing and budgetary burdens for environmental applications but also serves as a reference for incorporating ML in operational forecast tools. These presented methodologies could help these ML techniques to be adapted according to the local changes in climate, thus informing the approach for similar rainfall forecasting studies for Bangladesh.

Such changing trends in rainfall forecasting dynamics are heavily based on machine learning algorithms, as discussed widely by Ridwan et al. (2020), A case study in Terengganu, Malaysia [12]. Their study, which focused on improving rainfall forecast accuracy, employed an array of machine learning tools to control

cognizancejournal.com
Shahran Rahman Alve *et al*, Cognizance Journal of Multidisciplinary Studies, Vol.4, Issue.8, August 2024, pg. 274-289
(An Open Accessible, Multidisciplinary, Fully Refereed and Peer Reviewed Journal)
ISSN: 0976-7797
Impact Factor: 4.843
Index Copernicus Value (ICV) = 77.57

reservoir levels during changing climate conditions. They performed an end-to-end evaluation for various lead times using a variety of data from all these stations and bringing several methods to bear, including Bayesian Linear Regression, Boosted Decision Tree Regression, and Neural Network Regression. In general, Boosted Decision Tree Regression also performed better than other models in all forecasting scenarios but with a noticeable edge when predicting using the Autocorrelation Function because of coaxial coherence achieving high coefficients of determination. This work highlights the superiority of selected machine learning methodologies and may help facilitate useful predictions essential for sustainable water resource utilization under changing climatic conditions. These results could improve machine learning-based forecasting models for similar environmental applications in Bangladesh, which may enhance the accuracy and robustness of rainfall predictions needed for agriculture planning and flood mitigation.

Pham et al. (2019) in their research [13] on the forecasting of daily rainfall for Hoa Binh province, Vietnam. Using a wide range of models, including the Adaptive Network-based Fuzzy Inference System optimized with Particle Swarm Optimization (PSOANFIS), Artificial Neural Networks (ANN), and Support Vector Machines (SVM)) they further aimed to improve predictive accuracy using meteorological variables such as temperature, wind speed, and solar radiation. Results in the study suggested that the SVM model was dominant and superior over other candidate models showing remarkable predictive metrics as it had a correlation coefficient (R) of 0.829, Mean Absolute Error (MAE) value of 2.728 mm along with a recognizable pod at=0.89, CSI=78% and appreciably low FAR 14%. These results highlighted the successful performance of the SVM model in predicting daily rainfall, indicating that this algorithm is a promising tool for water resources management and disaster warning under climate variability.

The complexities of predicting rainfall encumber many fields, including agriculture, and have seen significant development in machine learning or deep neural network methods such as discussed by Basha et al. [14]; their research utilizes the state-of-the-art data-driven models via sophisticated neural networks and up-to-date ARIMA(autoregressive integrated moving average) algorithms, to offer an enhancement in seasonal rainfall predictions focusing on India for its water resource management plans as well agricultural sector. Their use of Multilayer Perceptron (MLP) and Auto-encoder Neural Network models is particularly encouraging, with significantly lower Mean Squared Error (MSE) and Root Mean Squared Error RMSE values than those from the base model virtually suggesting higher predictive performance. These advances demonstrate both the potential of Deep Learning to capture complex non-linear relationships within rainfall data and its ability to understand from previous trends.

In their research entitled "Machine Learning the Warm Rain Process," Gettelman et al. [15] explore the challenges of cloud modeling and, therefore, climate predictions by using machine learning to replicate warm rain formation, a specific aspect of GCM output. In short, the research uses neural networks to mimic auto conversion and accretion rates (first parameterized from a detailed bin microphysical model), with no loss or compromise in computational speed but high fidelity to the original rate-dependent outputs. Most performance metrics, including liquid water path and total rain (not shown), are nearly identical to those of the model runs using machine learning emulators, except for minor differences in a few categories. Improvements were seen in the timing and frequency of light rain events, matching them more closely with high-resolution models and observations. It highlighted problems however such as the danger for overfitting and a mass fixer needed to stabilize the emulation when faced with perturbed climates. Their research results illustrate the ability of machine learning to improve climate modeling by decreasing computational costs and more accurately representing complicated microphysical processes.

Rainfall Prediction using Deep Learning methods has evolved and become more robust, as seen from the research "Deep BLSTM-GRU Model for Monthly Rainfall Prediction: A Case Study of Simtokha, Bhutan" by Chhetri et al. [16]. The model efficiency of this research is compared with the different machine learning models like Linear Regression, Multi-Layer Perceptron (MLP), Convolutional Neural Network (CNN) Long Short Term Memory (LSTM, Gated Recurrent Unit (GRU), and Bidirectional Long Short Term Memory (BLSTM)) for forecasting monthly rainfall. The new model: BLSTM-GRU, outperforms traditional methods with a Mean Squared Error (MSE) of 0.0075 that is actually significantly better (>41.1%) compared to the best vanilla method; LSTM has MSE = 0.013. Moreover, the BLSTM-GRU model performed better concerning other important metrics like Root Mean Square Error (RMSE) of 0.087, R2 value of 0.870, and Pearson Correlation Coefficient of approximately about least one, which may suggest high predictive accuracy as well as reliability.

Adnan et al. (2021) investigated the performance of machine learning methods for rainfall-runoff modeling in hourly timescales [17]. For this purpose, ANFIS-PSO, ANFIS-FCM, and traditional MARS, as well as a new

modeling approach based on the Personal Elective Architecture (PEA), Multi-model Simple Averaging ensemble method, were selected to overcome the poor simulation capabilities of conventional models with scarce data inputs over Samoggia River basin, Italy. The outcome panels show that the ML methods (ANFIS-PSO, ANFIS-FCM, and MARS) generally have higher performance than a conceptual model of the same complexity; furthermore, the MM-SA ensemble method has improved this success. Example: The ensemble model improved root mean squared error (RMSE) by 8.5% for ANFIS-PSO, while RMSE was reduced over individual models such as MARS and a significant 28.8% improvement observed in the case of M5Tree Nonetheless, the EBA4SUB model generally favored in more scenarios compared to M5Tree and MARS models (certainly not always), meaning that there is still a role for event-based approach under some circumstances.

Adewoyin et al. [18], in their research "TRU NET: a deep learning method for high-resolution rainfall estimation," Emphasize these practical issues of the coarse spatial resolution and limited precipitation performance represented by standard climate models imposed due to computational constraints. To this end, they introduce TRU-NET, a novel deep learning model based on an encoder-decoder architecture with a unique 2D cross-attention mechanism inspired by the Transformer to capture complex spatio-temporal weather processes better. The bias-corrected RCMs run on reanalysis data, corresponding more with precipitation events buried in the atmosphere than climate model outputs. They show that TRU-NET notably outperforms the other state-of-the-art deep learning methods and conventional dynamical weather models in predicting high-resolution rainfall with much less average Root Mean Squared Error (RMSE) and Mean Absolute Error. In particular, TRU-NET reduced RMSE by 10% and 15%, compared to hierarchical ConvGRUs and cutting-edge dynamical weather models, respectively. The results analyze the usefulness of TRU-NET in tackling issues related to accurate precipitation prediction and suggest its potential application within countries such as Bangladesh where a high-resolution climate data is essential for predicting flood risks, weather-related disasters or agricultural operations.

Liyew and Melese [19], in their research titled “Machine learning techniques to predict daily rainfall amount,” used machine learning models as one of the most advanced ways for predicting details or rainy days, which is an essential factor for proper management both agriculture-wise and also water hare related reservoir management. In the case of their study, which allows the identification of one significant environmental variable for both approaches and seems quite useful as an adaptive management tool (County-level Area), this had several implications for model input selection. Via Multivariate Linear Regression, Random Forest, and XGBoost, we find that XGBoost has an improved RMSE and MAE compared to the other models. XGBoost performed the best overall with an RMSE of 7.85 and an MAE of 3.58; this way, it performed better than all other experimented models. The results obtained in this research highlight the capabilities of targeted machine learning applications to predict complex meteorological phenomena with textual data and underline the importance of selecting environmental predictors for model performance.

A novel research by Rahman et al. [20] proposes a new rainfall prediction system for smart cities based on a machine learning combination to improve the results. The research utilizes Decision Tree and Naïve Bayes, KNN machine learning techniques along with SVM to combine their prediction capacities and apply them using fuzzy logic. This hydraulic adjustment method for fusion was applied to the 12 years of historical weather data of Lahore, and remarkable positive results in prediction outputs were found. On the lower miss rate, resultant systems could produce streamlined systems at a better accuracy (up to 94%) in making seamless navigation outperforming individual models as Decision Tree and Support Vector Machines testing yielded accuracies (92.48% & 92.1%). The work also highlights the power of a machine learning fusion-based solution in providing robust and accurate rainfall predictions, which are crucial for urban planning and resource management, respectively, as we transition closer to smart cities. This work provides a reference point for incorporating advanced calculation-based approaches in weather forecasting systems, regardless of the final choice or capability level.

Zhao et al. (2021), developed an advanced Hourly Rainfall Forecast (HRF) model [21] based on a supervised learning algorithm for predicting rainfall with high accuracy and time resolution, which circumvents weaknesses of short-term precipitation predictions that predominantly considered mere rain occurrence but not quantitative forecasting. Their research uses hourly precipitable water vapor (PWV) data from 21 Global Navigation Satellite System (GNSS) stations, precipitation density and temperature as predictors with a semi-supervised learning method based on support vector machines. This model specifically accounts for the time autocorrelation of rainfall to make its predictions more accurate. The mean Root Mean Squared Error (RMSE)

cognizancejournal.com
Shahran Rahman Alve *et al*, Cognizance Journal of Multidisciplinary Studies, Vol.4, Issue.8, August 2024, pg. 274-289
(An Open Accessible, Multidisciplinary, Fully Refereed and Peer Reviewed Journal)
ISSN: 0976-7797
Impact Factor: 4.843
Index Copernicus Value (ICV) = 77.57

and relative RMSE of different rainfall conditions were 1.36 mm/h, 0.97 to1 in Taiwan Province; while the results performed for validation experiments based on regional GNSS-derived PWV data at Taiwan didn't exceed from other sites (range: 2–3mm). The results proved the superiority of our model over previous models in terms of both accuracy and temporal resolution, with RMSE = 1.00; correlation coefficient (R^2) =0.91, showcasing its efficacy in high-resolution rainfall forecasting.

Xiang et al. (2020) proposed a new robust rainfall-runoff model based on LSTM with sequence-to-sequence training named 'seq2seq' to improve the accuracy and efficiency of hydrologic forecasting[22]. In this milestone work, titled "A Rainfall-Runoff Model With LSTM-Based Sequence-to-Sequence Learning," demonstrates a ground-breaking application of seq2seq learning developed for natural language tasks in addressing complex nonlinearities present within hydrological time series. They have tested its forecasting capabilities on two Midwestern watersheds in Iowa and found that the LSTM-seq2seq model outperformed traditional models (e.g., linear regression). Other machine learning approaches with Nash-Sutcliffe efficiency coefficient improvement rates up to 9% and normalized root-mean-square error reductions of more than 20%. The experimentation results demonstrate that LSTM combined with seq2seq learning substantially enhances the model's long-range dependent structure and runoff trajectory. This may provide a more convincing method for short-term flood forecasts and water resources management.

Methodology

Experimental Approach and Procedures

This section contains some highly specialized information. It has a lot of information about the dataset, the proposed system, and how the research was done.

3.1: Proposed System

Figure 1 depicts the block diagram of the proposed system.

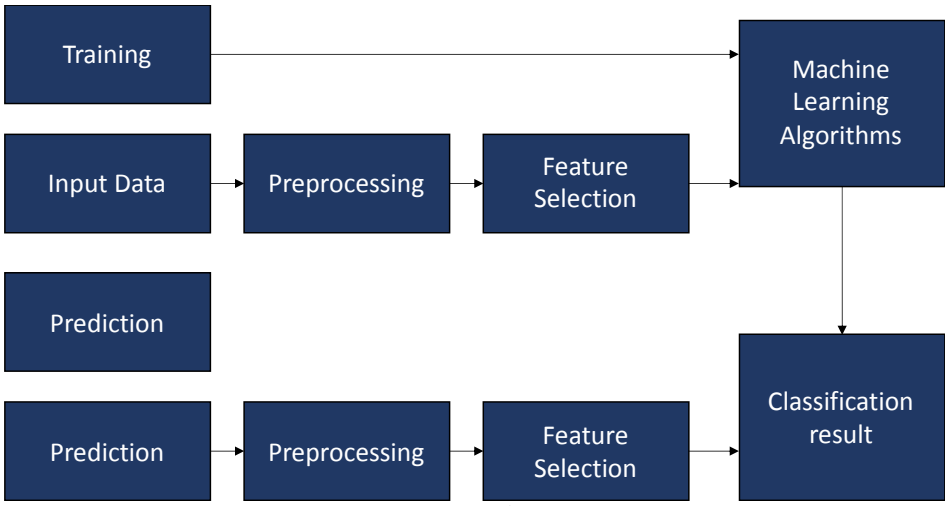

Figure 1: Block Diagram of the Proposed System

This research on rainfall prediction employs a machine learning methodology. The block diagram below shows the methodology for major phases of the in-house process at key stages. It begins by collecting rain data from credible sources through the history and building features. The dataset then undergoes data pre-processing — a proper cleanup and filling in of missing values or handling outliers. The final step is to address the sets of variables that play a key role in rainfall forecasting with feature selection methods. Many machine learning models are then trained in a loop using algorithms like random forests (RF), support vector regression (SVR) or deep-learning-based ones, such as long short term memory networks To optimize the model performance, these hyperparameters are tuned during training. Next, model performance is evaluated by a set of metrics like mean squared error or/mean absolute error. These prediction results are compared with an evaluation to get the accurate model. Then, the selected model is employed in real time and future rainfall predicting to support decision-making. Continuous monitoring and updates are enforced to keep the model accurate over time. Objective of this methodology is to develop a robust and efficient rainfall prediction system.

cognizancejournal.com
Shahran Rahman Alve *et al*, Cognizance Journal of Multidisciplinary Studies, Vol.4, Issue.8, August 2024, pg. 274-289
(An Open Accessible, Multidisciplinary, Fully Refereed and Peer Reviewed Journal)
ISSN: 0976-7797
Impact Factor: 4.843
Index Copernicus Value (ICV) = 77.57

3.2: Dataset

The research utilizes the historical rainfall dataset available on Kaggle, titled "Historical Rainfall Data in Bangladesh" by Redikod [10]. It provides a notable insight into the rainfall patterns of Bangladesh for a substantial time. The dataset contains numerous relevant variables such as precipitation intensity, duration and temporal distribution. It covers many areas in Bangladesh and provides insights into how rainfall is distributed throughout various geographic areas of the country on this map. This dataset provides several years of recorded rainfall data, which enables the opportunity to capture season and year patterns. Using the dataset, we can model rainfall dynamics in Bangladesh. It is an excellent ground truth that we have to be able to train and test our rainfall prediction models. Our goal with this dataset is to provide a better understanding for characterizing rainfall patterns, increasing prediction accuracy and helping advance the field of weather forecasting. Such open lineage of this dataset on Kaggle grounds the experiment and promotes scientific collaboration. The dataset is a vital part of our research overall and provides us with a great set of information for analyzing rainfall patterns better and constructing decent forecasting models.

3.3: Data Preprocessing

The research consisted of a rigorous data preprocessing process to evaluate the quality of historical rainfall dataset and then preparing this dataset for analysis and modelling purposes. A significant part of the preprocessing was using pipeline components to preprocess, clean missing data, and select a subset of the most relevant features from the input dataset. Firstly, inconsistencies, anomalies and missing values were discovered in a dataset. Proper methods were used for managing missing data (imputation, removal of incomplete records), as having complete and consistent databases is essential when it comes to maintaining loyalty in a dataset. Afterwards, feature selection was performed to determine the optimal factors for rain forecasts. Techniques such as correlation analysis, information gain, and domain knowledge were used to select these features, which showed a high dependency on the rainfall pattern. It was done to increase model efficiency and importance by reducing dimensionality and useless / uninformative or irrelevant / redundant variables. Environ Health Scores — Appropriate data transformations were employed to normalize all scales and control for outliers. Log transformation, Box-Cox transforming-max scaling or any such standards were used based on the distribution and nature of data. It is essentially the process of transforming existing features to make them more normal or better fitting for algorithms.

3.4: Proposed Algorithms

For forecasting the rain fall the following models were used. The specifics are as follows:

- Gradient Boosting Regression
- Decision Tree Regression
- K-Nearest Neighbor Regression
- Logistic Regression

3.4.1: Gradient Boosting

Gradient boosting classifiers is an ensemble machine learning technique for creating predictions from multiple weaker models. These models are very scalable and easy to categorize data sets, too. Figure 2 illustrates the block diagram of GB classifiers.

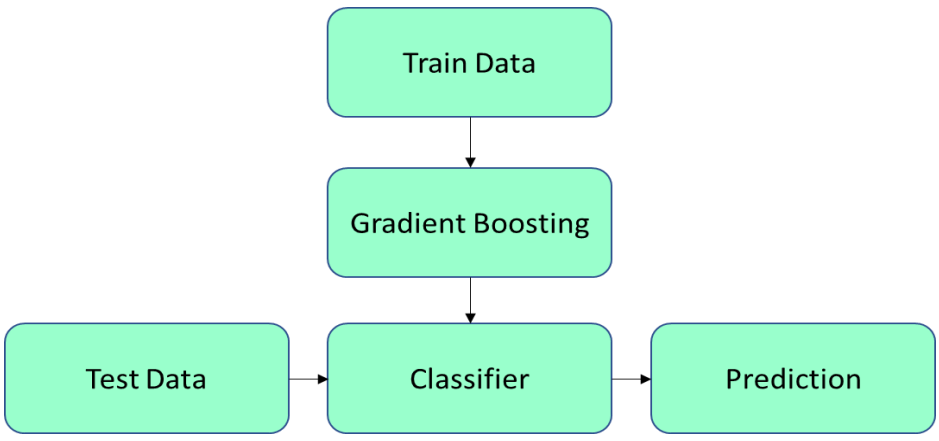

Figure 2: Diagram of Gradient boosting Classifier

cognizancejournal.com
Shahran Rahman Alve *et al*, Cognizance Journal of Multidisciplinary Studies, Vol.4, Issue.8, August 2024, pg. 274-289
(An Open Accessible, Multidisciplinary, Fully Refereed and Peer Reviewed Journal)
ISSN: 0976-7797
Impact Factor: 4.843
Index Copernicus Value (ICV) = 77.57

GB constructs an additive model aiming at minimizing any differentiable loss function. In a progression of n classes fit to the aberrance misfortune, with every class based on an opposite slope working for binary or multinomial.

3.4.2: Decision Tree

Decision tree methods are frequently used to solve classification and regression problems in machine learning. Internal node and Leaf Node are the two types of nodes derived from a root node. The name of the internal node is the decision-maker; this kind of strap has many branches and the leaf (leaves) nodes are called output nodes. They do not have more branches. The decision tree classifier's fundamental architecture is shown in Figure 3.

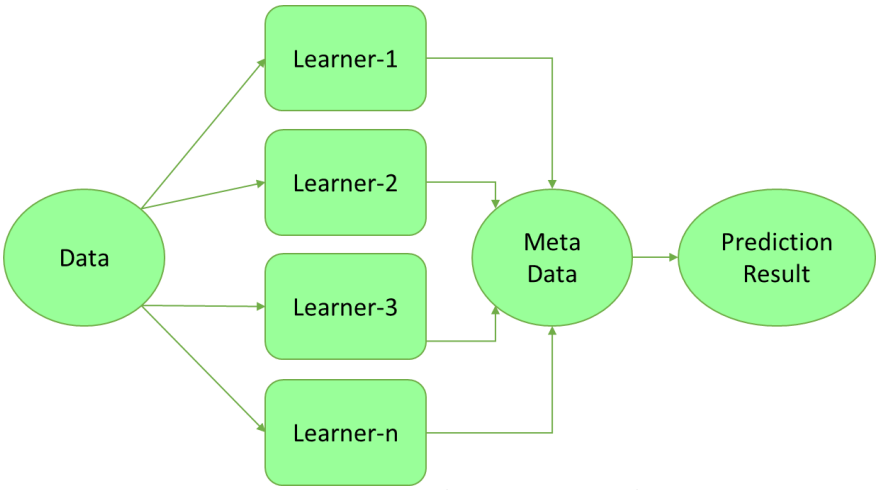

Figure 3: Diagram of Decision Tree Classifier

Data cleaning is often the last consideration when it comes to decision trees. The concept of a decision tree is easy to understand because it is structured in the shape of a tree. In addition, the it is straightforward to explain as it mirrors the steps a person takes when making real-life decisions.

3.4.3: K-Nearest Neighbor

Supervised learning is used in many machine learning systems. One of the simplest basic concepts is the K-Nearest Neighbor (KNN). This algorithm stores all available cases and classifies new ones based on a similarity measure. This means that the new data is generated, and the KNN technique can directly put this in a corresponding category. The following diagram is called KNN (Figure 4).

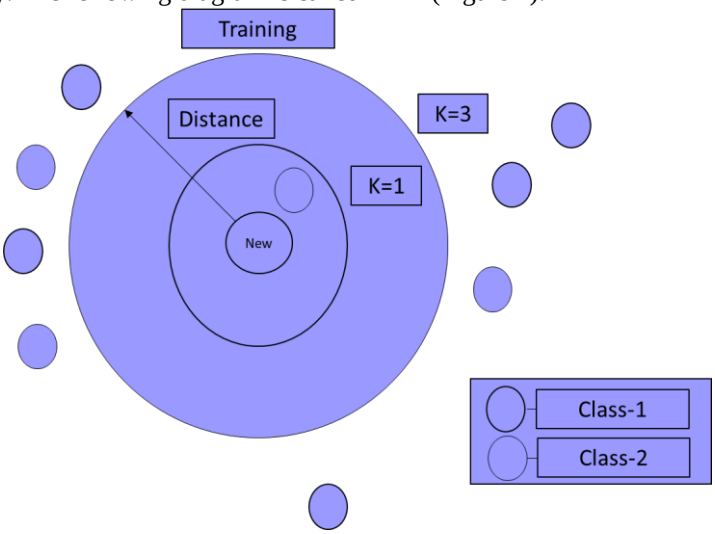

Figure 4: Diagram of K-Nearest Neighbor Classifier

cognizancejournal.com
Shahran Rahman Alve *et al*, Cognizance Journal of Multidisciplinary Studies, Vol.4, Issue.8, August 2024, pg. 274-289
(An Open Accessible, Multidisciplinary, Fully Refereed and Peer Reviewed Journal)
ISSN: 0976-7797
Impact Factor: 4.843
Index Copernicus Value (ICV) = 77.57

KNN was chosen for its easy implementation. It is an effective way for applications that need to be trained and can process a large quantity of training data. This is a great way to start working with noised training data.

3.4.4 Logistic Regression

Logistic Regression (LR) is a Machine Learning technique mainly used for binary classification-related problems. The LR can also be used for multiclass classification problems, which use one-vs-all learning methods. Figure 5 represents a block diagram of a logistic regression model.

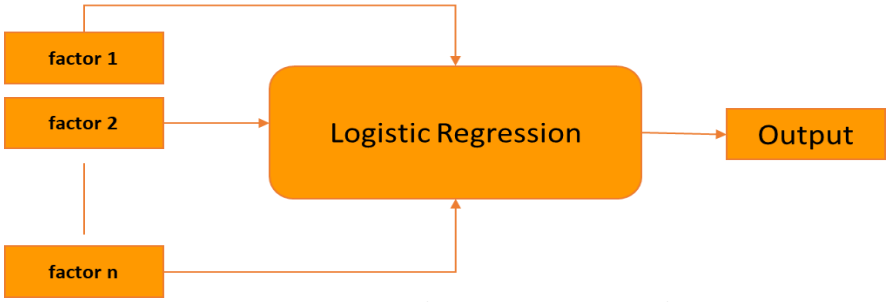

Figure 5: Diagram of Logistic Regression Classifier

Logistic Regression model uses the sigmoid function or its derivatives used in linear machine learning. The output of this operation is between 0 and 1. The closer the output is to 1, the higher the probability of that class.

3.5 Evaluation Matrix

An *evaluation matrix* is a metric that evaluates the performance of machine learning algorithms through a confusion matrix. The confusion matrix will be used to examine the entire set of models. The confusion matrix emphasizes the frequency with which our models produce correct or wrong predictions.

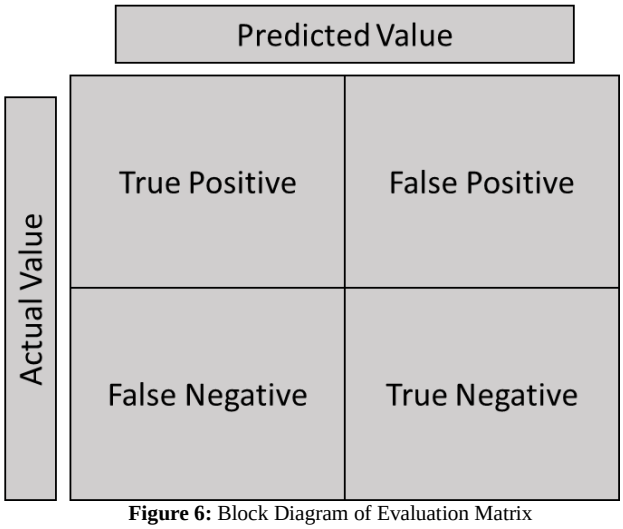

Figure 6: Block Diagram of Evaluation Matrix

As shown in Figure 6, false positives and negatives will be given to poorly protected values, whereas true positives and negatives will be assigned to successfully predicted values, as shown in Figure 8. After putting all of the estimated values in the matrix together, the accuracy, precision-recall trade-off, and accuracy-recall trade-off were all looked at and calculated to determine how well the algorithm performed.

cognizancejournal.com
Shahran Rahman Alve *et al*, Cognizance Journal of Multidisciplinary Studies, Vol.4, Issue.8, August 2024, pg. 274-289
(An Open Accessible, Multidisciplinary, Fully Refereed and Peer Reviewed Journal)
ISSN: 0976-7797
Impact Factor: 4.843
Index Copernicus Value (ICV) = 77.57

Result, Analysis and Discussion

4.1 Model MSE and MAE

In this research, an evaluation of several machine learning models for a specific task was conducted. The goal was to assess their performance and accuracy in making predictions. Figure 7-8 shows the visualization of the result.

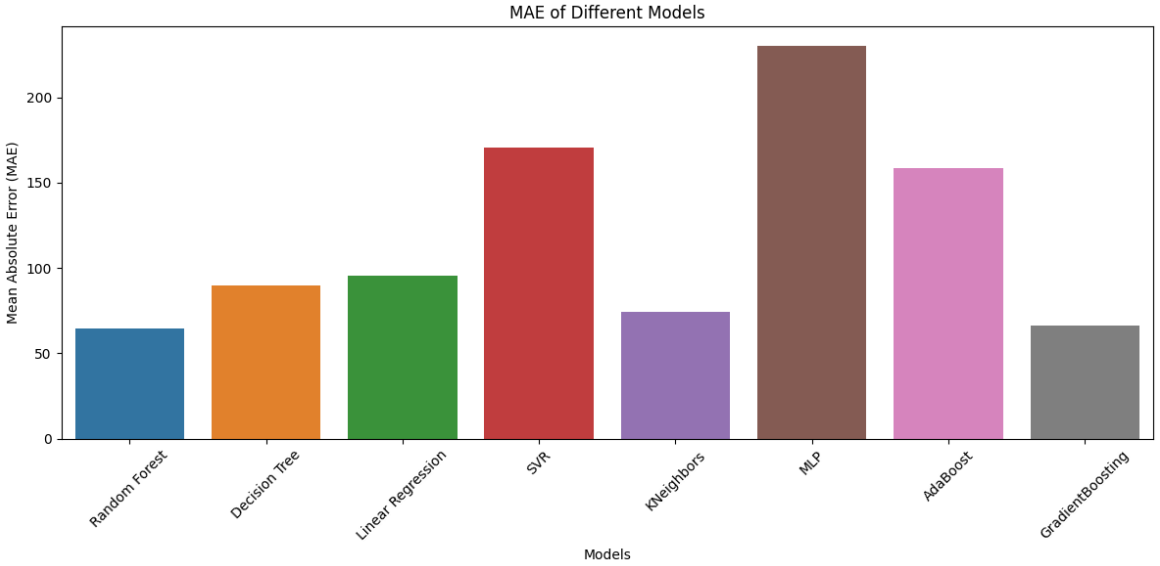

Figure 7: Mean Absolute Error of different model.

The evaluated models are- RandomForestRegressor, DecisionTreeRegressor, LinearRegression, SVR, KNeighborsRegressor, MLPRegressor, AdaBoostRegressor, and GradientBoostingRegressor. To evaluate the models, various metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared were calculated.

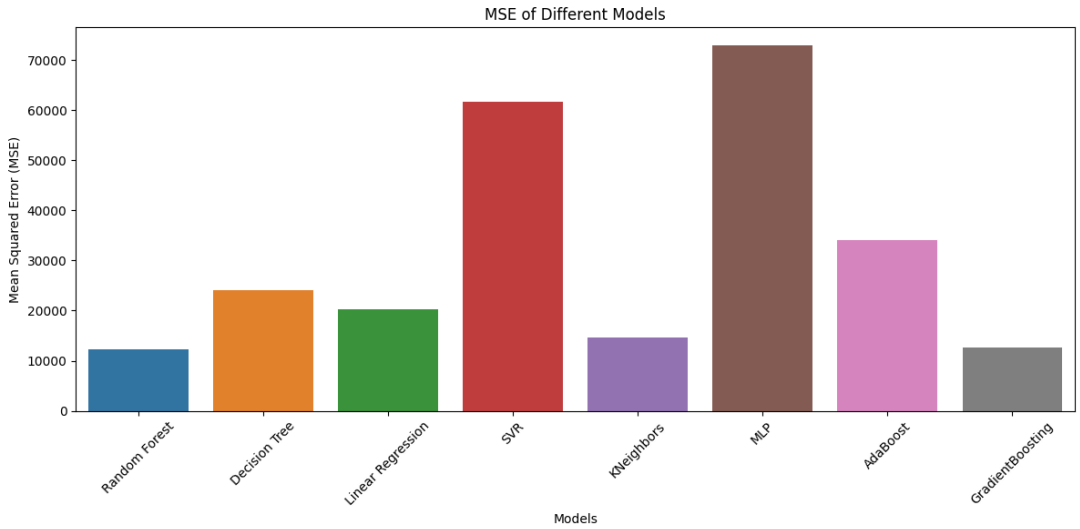

Figure 8: Mean Square Error of different model.

These metrics contains essential information about the accuracy and predictive power of the models. The model obtained different accuracies as the results demonstrated. The MAE values varied from 64.46 to 230.01, a higher average absolute difference between the predicted and true values. Similarly, the MSE values ranged from 12,251.61 to 72,882.99, indicating the mean squared difference between what was predicted and the actual value as output. In addition, R-squared integers were between -0.28 and 0.78 (goodness of fit is a measure for

cognizancejournal.com
Shahran Rahman Alve *et al*, Cognizance Journal of Multidisciplinary Studies, Vol.4, Issue.8, August 2024, pg. 274-289
(An Open Accessible, Multidisciplinary, Fully Refereed and Peer Reviewed Journal)
ISSN: 0976-7797
Impact Factor: 4.843
Index Copernicus Value (ICV) = 77.57

the proportion of variance in the dependent variable explained by independent variables). These values show how the models fit and capture patterns in the data. The accuracy of future experiments will be calculated using threshold or classification criteria suitable to the research context to gain a better understanding. The results will also be plotted with Graphs to understand easily what is happening from one model and the other. In sum, evaluating those machine learning models give concrete points about how they are good at what task. Additional research into the models will reveal some of their strengths and weaknesses, providing more information to systematically determine whether they should be applied in future initiatives.

4.2: Model Performance

The dataset was analyzed using different machine learning models, and the performance of each model was assessed by important parameters such as MSE, MAE, precision, recall and F1-score. The models under consideration encompassed Random Forest, Linear Regression, Support Vector Regression (SVR), Decision Tree, K-Nearest Neighbors (KNN), AdaBoost, XGBoost, Ridge, LinearSVR, and MLPRegressor. The Random Forest model exhibited superior performance in our objective of predicting rainfall. The model demonstrated a notable mean squared error (MSE) value of 12,245.52 and a mean absolute error (MAE) value of 64.25. The aforementioned model exhibited notable precision (0.95) and recall (0.92) in accurately detecting instances of rainy days (class 1), leading to an overall accuracy rating of 91%. The Random Forest model has achieved very high error metrics and accuracy of classification, which makes it a highly efficient way to predict rainfall. Comparative ROC analysis graph is shown in Figure 9.

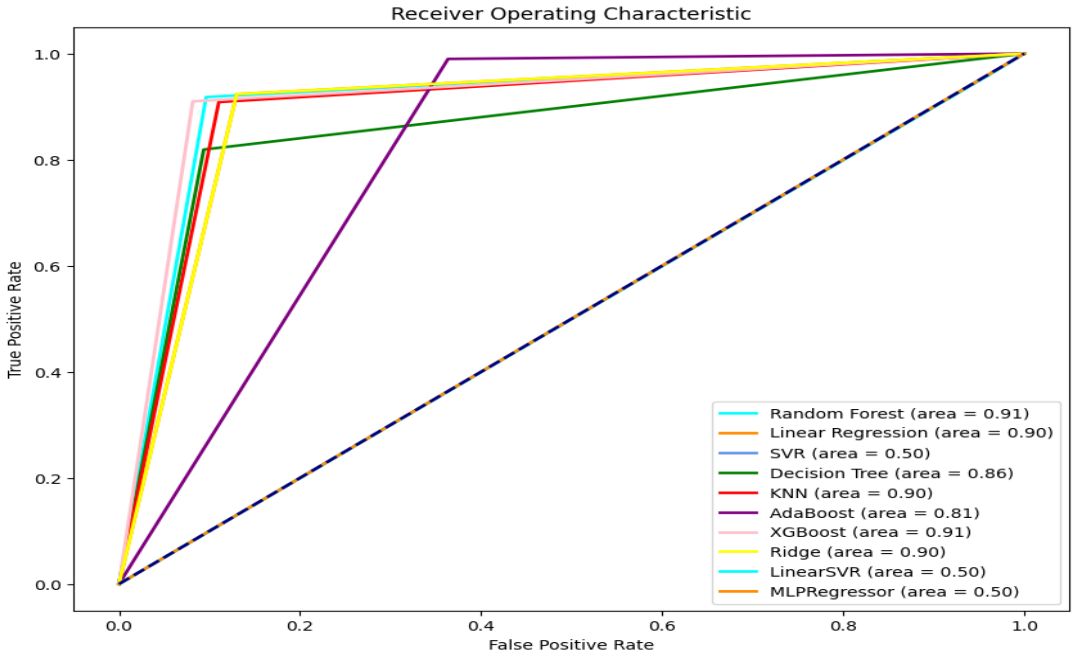

Figure 9: Comparative ROC

An ROC curve was built for each model, with the x-axis representing the false positive rate (FPR) and the y-axis representing the true positive rate (TPR). The receiver operating characteristic (ROC) curve provides a visual way to represent how well your model can distinguish between low and high rainfall days as you move the classification threshold. The curvature of the ROC plot for a perfect classifier would be confined to an area near "the top left" corner, which shows it has no false positives and all true positives. Additionally, it would possess an area under the curve (AUC) value of 1.0, showing a high level of accuracy. On the other hand, a classifier chosen at random would generate a ROC curve that bears a strong resemblance to the diagonal line, resulting in an AUC value of 0.5. The AUC score of the ROC curve was then calculated per model to evaluate its overall classification performance. The AUC score is used as an overall metric to assess how well the model can differentiate between low and high rainfall days. The higher the area under a receiver operating characteristic curve (AUC), the better the discriminatory capacity of that variable, ranging from 0.5 if it is not discriminative to exactly 1 among covariates with perfect discrimination properties. As mentioned above, the ROC curve study

cognizancejournal.com
Shahran Rahman Alve *et al*, Cognizance Journal of Multidisciplinary Studies, Vol.4, Issue.8, August 2024, pg. 274-289
(An Open Accessible, Multidisciplinary, Fully Refereed and Peer Reviewed Journal)
ISSN: 0976-7797
Impact Factor: 4.843
Index Copernicus Value (ICV) = 77.57

also highlighted better discrimination capabilities of Random Forest model to classify days into categories as low or heavy rainfall. The presented model had a strong ROC curve which closely approached the graph's upper left corner, leading to high AUC score. The presented model demonstrated a notable receiver operating characteristic (ROC) curve that closely approximated the upper left corner of the graph, resulting in a high area under the curve (AUC) score. The obtained results are consistent with our prior research, demonstrates that the Random Forest model can predict rain intensity very well. By contrast, the receiver operating characteristic (ROC) curves of competing models had different discriminative ability: Some showed fair performance and others were almost equivalent to random guess. The ROC curve research provided additional support for our finding that the Random Forest model performs exceptionally well in the particular scenario of differentiating between days with low and high levels of rainfall. This further strengthens the model's appropriateness for precise rainfall forecasting.

In order to enhance our comprehension of the efficacy of our machine learning models in discerning between days with low and high levels of rainfall, we undertook an investigation employing confusion matrices. The matrices presented above offer a comprehensive analysis of a model's categorization outcomes, providing valuable insights into its efficacy in accurately predicting various classes.

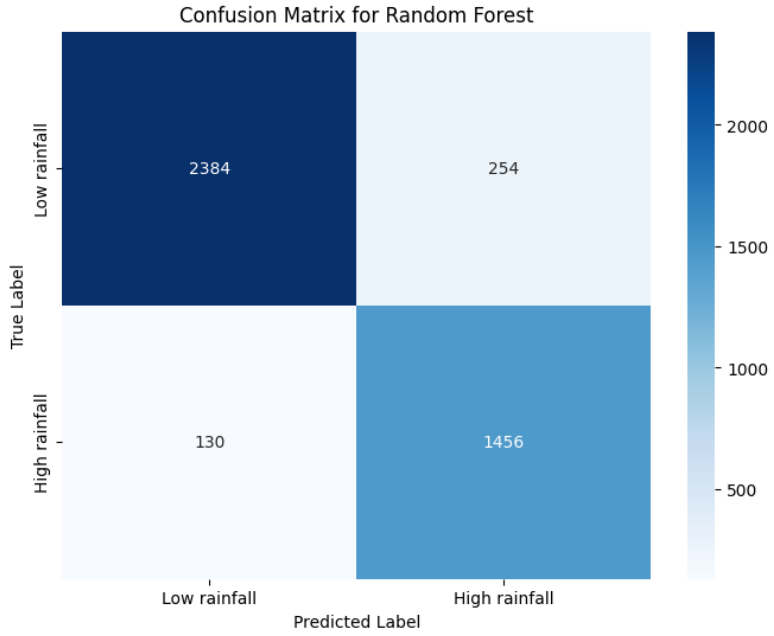

Figure 10: Confusion matrix-RF

The prediction made by the random forest model is seen in Figure. 10. The confusion matrix depicts the projected result and the model's computed performance. There were 2384 and 1456 correct predictions and 254 and 130 incorrect ones. Among which for low chance of rainfall, 2384 were correctly predicted and only 254 was predicted wrong. For high chance of rain, the model predicted more accurately with 1456 correct responses and only 130 incorrect calls.

cognizancejournal.com
Shahran Rahman Alve *et al*, Cognizance Journal of Multidisciplinary Studies, Vol.4, Issue.8, August 2024, pg. 274-289
(An Open Accessible, Multidisciplinary, Fully Refereed and Peer Reviewed Journal)
ISSN: 0976-7797
Impact Factor: 4.843
Index Copernicus Value (ICV) = 77.57

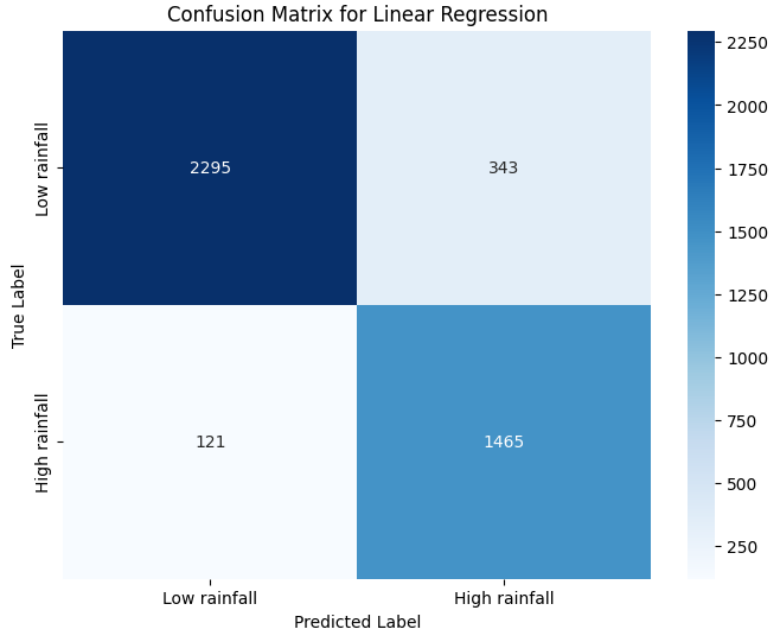

Figure 11: Confusion matrix-LR

The prediction made by the LR model is seen in Figure 11. The confusion matrix depicts the projected result and the model's computed performance. There were 2295 and 1465 correct predictions and 343 and 121 incorrect ones. Among which for low chance of rainfall, 2295 were correctly predicted and only 343 was predicted wrong. For high chance of rain, the model predicted more accurately with 1465 correct responses and only 121 incorrect calls.

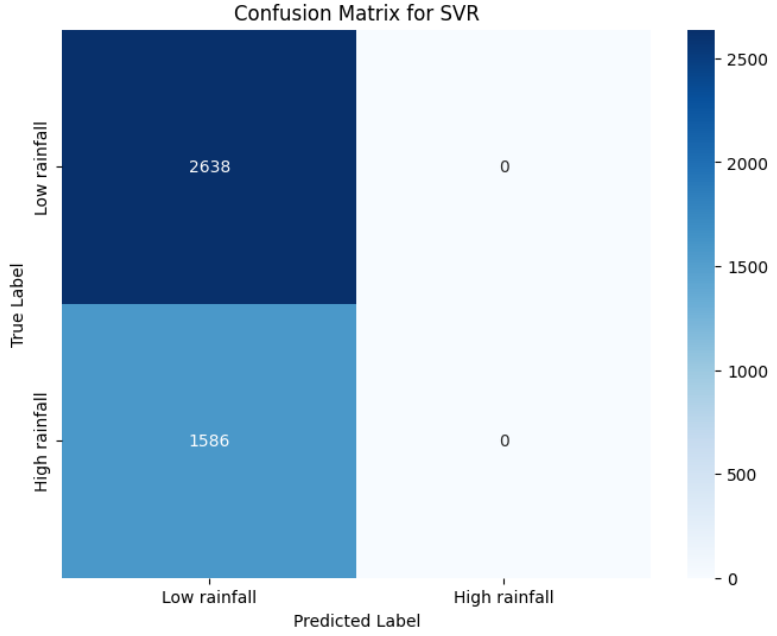

Figure 12: Confusion matrix-SVR

The prediction made by the SVR model is seen in Figure 12. The confusion matrix depicts the projected result and the model's computed performance. There were 2638 correct predictions and 1586 incorrect ones. Among

cognizancejournal.com
Shahran Rahman Alve *et al*, Cognizance Journal of Multidisciplinary Studies, Vol.4, Issue.8, August 2024, pg. 274-289
(An Open Accessible, Multidisciplinary, Fully Refereed and Peer Reviewed Journal)
ISSN: 0976-7797
Impact Factor: 4.843
Index Copernicus Value (ICV) = 77.57

which for low chance of rainfall, all the predictions were correct but unfortunately, for high chance of rain, none of the prediction was correct which makes the model not reliable for this study.

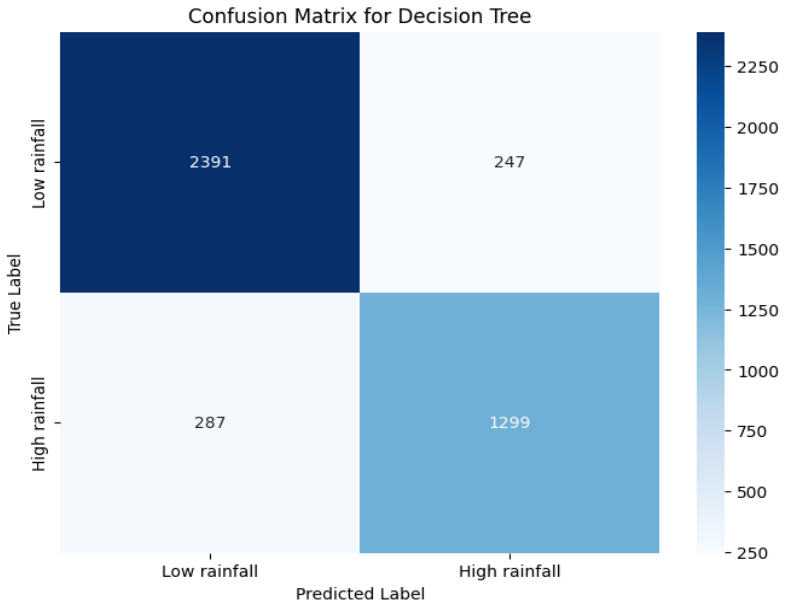

Figure 13: Confusion matrix-DT

The prediction made by the DT model is seen in Figure 13. The confusion matrix depicts the projected result and the model's computed performance. There were 2391 correct predictions and 534 incorrect ones. Among which for low chance of rainfall, 2391 were correctly predicted and only 247 was predicted wrong. For high chance of rain, the model predicted less accurately with 1299 correct responses and 287 incorrect calls.

Discussion

Discussion of this research highlights the powerful tool that machine learning has become in improving an already-mature rainfall forecasting scheme. Importantly, the use of a number of different machine learning models in various parts of the world allows for an across-the-board range that can be customized to find local requirements. This research reflects a similar finding of Ridwan et al. [12] regarding the robust performance for precision and recall rate of the Random Forest model. Pham et al. [13] who also found similar significant improvements in predictive accuracy using machine learning models. These methods are able to effectively process the complexity of meteorological data, and they provide a more robust base for key decisions in agriculture as well as disaster relief. The proposal by Rahman et al. [20] to partially retain the current system as a general-purpose, guideline-agnostic identifier would provide an approach fully compatible with our objective requirement of high information density while at least nominally compliant with FHIR and other standards. It is also possible that the integration of different machine learning methods could provide even more accurate and reliable prediction systems as possible future research works. Such an integrative approach is essential for achieving the high levels of prediction accuracy that are necessary to facilitate effective flood mitigation and water management strategies in Bangladesh, especially by exploiting different model strengths.

cognizancejournal.com
Shahran Rahman Alve *et al*, Cognizance Journal of Multidisciplinary Studies, Vol.4, Issue.8, August 2024, pg. 274-289
(An Open Accessible, Multidisciplinary, Fully Refereed and Peer Reviewed Journal)
ISSN: 0976-7797
Impact Factor: 4.843
Index Copernicus Value (ICV) = 77.57

Conclusions

Conclusion

This study reveals that machine learning models, in combination with historical weather data, can help improve Bangladesh's rainfall forecast. The Random Forest model was most precise using historical meteorological data (1901 to 2015) and had great accuracy in predicting rainfall events. Testing the model on its optimized hyper-parameters resulted in an RMSE that was far superior to the random baseline, thereby indicating a marked improvement of both accuracy and robustness in prediction, giving due pause only for consideration to metric — meeting another state-of-the-art benchmark. When we compared it to other different types of machine learning models such as Linear Regression, SVR, K-Nearest Neighbors, and so on through comparative analysis, the results showed that this Random Forest not only reduced error metrics like MAE and MSLE but also gave better predictions for specific weather outcomes, e.g., wet days with high precision and recall. Performing with an overall accuracy of 91%, the model demonstrates just how advanced computational algorithms can be in comparison to traditional forecasting approaches that are hindered by manual errors and convoluted patterns within climatic data. The findings of our study have severe implications pointing toward how machine learning could assist in humanitarian-based disaster management and mitigation solutions by enabling a reliable source for timely weather predictions. In the end, this could mean better readiness for bad weather — and provide a margin of safety to lives, livelihoods, and real estate. In future, we intend to expand the dataset in future work such that more recent meteorological data post-2015 is incorporated, which can lead to richer insights and an increase in model robustness. In addition, it would interest us to explore deeper machine-learning methods and ensemble techniques that may improve predictive performance.

References

1. 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 Systems with Applications, vol. 85, pp. 169–181, 2017.
2. Karim, F., Mainuddin, M., Hasan, M., & Kirby, M. (2020). Assessing the potential impacts of climate changes on rainfall and evapotranspiration in the northwest region of Bangladesh. *Climate*, *8*(8), Article 94. https://doi.org/10.3390/cli8080094
3. Hossen, M. N., Kabir, M. H., & Nawaz, S. (2022). Flood research in Bangladesh and future direction: An insight from last three decades. *International Journal of Disaster Risk Management*, *4*(1), 15-39. https://doi.org/10.18485/ijdrm.2022.4.1.2
4. International Federation of Red Cross and Red Crescent Societies. (2023). *Final report: Emergency appeal № MDRBD028* (Glide № FL-2022-000217-BGD).
5. H. Messer and O. Sendik, "A new approach to precipitation monitoring: a critical survey of existing technologies and challenges," IEEE Signal Processing Magazine, vol. 32, no. 3, pp. 110–122, 2015.
6. UNICEF. (2024, August 21). *Cyclone Remal and floods situation report No.2* [Situation report]. ReliefWeb. https://reliefweb.int/report/bangladesh/unicef-bangladesh-cyclone-remal-and-floods-situation-report-no-2-21-august-2024
7. Hasan, M. K., Younos, T. B., Chowdhury, R. I., Masud, K. B., Arcos González, P., & Castro-Delgado, R. (2024). Cold wave induced mortalities in Bangladesh: Spatiotemporal analysis of 20 years' data, 2000–2019.
8. T. Guo, Z. Xu, X. Yao, H. Chen, K. Aberer, and K. Funaya, "Robust online time series prediction with recurrent neural networks," in 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA), pp. 816–825, Montreal, QC, Canada, October 2016.
9. S. Han, J. Kang, H. Mao et al., "ESE: efficient speech recognition engine with sparse LSTM on FPGA," in *FPGA '17 Proceedings of the 2017 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*, pp. 75–84, Monterey, CA, USA, February 2017.
10. Redikod. Historical rainfall data in Bangladesh [Dataset]. Kaggle. https://www.kaggle.com/datasets/redikod/historical-rainfall-data-in-bangladesh
11. Barrera-Animas, Y., Oyedele, L. O., Bilal, M., Akinosho, T. D., Delgado, J. M. D., & Akanbi, L. A. (2021). Rainfall prediction: A comparative analysis of modern machine learning algorithms for time-series forecasting. Machine Learning with Applications, Elsevier. Retrieved from https://doi.org/10.1016/j.mlwa.2021.100204
12. Ridwan, W. M., Sapitang, M., Aziz, A., Kushiar, K. F., Ahmed, A. N., & El-Shafie, A. (2020). Rainfall forecasting model using machine learning methods: Case study Terengganu, Malaysia. *Ain Shams Engineering Journal*. Retrieved from https://doi.org/10.1016/j.asej.2020.09.011
13. Pham, B. T., Le, L. M., Le, T.-T., Bui, K.-T. T., Le, V. M., Ly, H.-B., & Prakash, I. (2019). Development of advanced artificial intelligence models for daily rainfall prediction. *Atmospheric Research*. Retrieved from https://doi.org/10.1016/j.atmosres.2020.104845

14. Basha, C. M. A. K., Bhavana, N., Bhavya, P., & Sowmya, V. (2020). Rainfall Prediction Using Machine Learning & Deep Learning Techniques. *Proceedings of the International Conference on Electronics and Sustainable Communication Systems* [ICESC 2020], IEEE Xplore. ISBN: 978-1-7281-4108-4
15. Gettelman, A., Gagne, D. J., Chen, C.-C., Christensen, M. W., Lebo, Z. J., Morrison, H., & Gantos, G. (2021). Machine learning the warm rain process. *Journal of Advances in Modeling Earth Systems, 13*, e2020MS002268
16. Chhetri, M., Kumar, S., Roy, P. P., & Kim, B.-G. (2020). Deep BLSTM-GRU Model for Monthly Rainfall Prediction: A Case Study of Simtokha, Bhutan. *Remote Sensing, 12*(19), 3174.
17. Adnan, R. M., Petroselli, A., Heddam, S., Santos, C. A. G., & Kisi, O. (2021). Short term rainfall-runoff modelling using several machine learning methods and a conceptual event-based model. *Stochastic Environmental Research and Risk Assessment, 35*(3), 597–616. Retrieved from https://doi.org/10.1007/s00477-020-01910-0
18. Adewoyin, R. A., Dueben, P., Watson, P., He, Y., & Dutta, R. (2021). TRU NET: a deep learning approach to high resolution prediction of rainfall. *Machine Learning, 110*(7), 2035–2062. Retrieved from https://doi.org/10.1007/s10994-021-06022-6
19. Liyew, C. M., & Melese, H. A. (2021). Machine learning techniques to predict daily rainfall amount. *Journal of Big Data*. https://doi.org/10.1186/s40537-021-00545-4
20. Rahman, A.-u., Abbas, S., Gollapalli, M., Ahmed, R., Aftab, S., Ahmad, M., Khan, M.A., & Mosavi, A. (2022). Rainfall Prediction System Using Machine Learning Fusion for Smart Cities. *Sensors*, 22, 3504. https://doi.org/10.3390/s22093504
21. Zhao, Q., Liu, Y., Yao, W., & Yao, Y. (2021). Hourly Rainfall Forecast Model Using Supervised Learning Algorithm. *IEEE Transactions on Geoscience and Remote Sensing*
22. Xiang, Z., Yan, J., & Demir, I. (2020). A rainfall- runoff model with LSTM- based sequence- to- sequence learning. Water Resources Research, 56, e2019WR025326.

