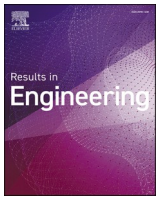
Results in Engineering 22 (2024) 102093

Contents lists available at ScienceDirect

Results in Engineering

journal homepage: www.sciencedirect.com/journal/results-in-engineering



Comparative analysis of different rainfall prediction models: A case study of Aligarh City, India

Mohd Usman Saeed Khan [a], Khan Mohammad Saifullah [b], Ajmal Hussain [a], Hazi Mohammad Azamathulla [c,*]

[a] Department of Civil Engineering, Zakir Hussain College of Engineering & Technology, Aligarh Muslim University, Aligarh, 202002, India
[b] Department of Industrial Chemistry, Faculty of Science, Aligarh Muslim University, Aligarh, 202002, India
[c] Department of Civil and Environmental Engineering, University of the West Indies, Saint Augustine, Trinidad and Tobago

ARTICLE INFO

Keywords:
Rainfall prediction
Deep learning
Machine learning
Water resources

ABSTRACT

This research paper delves into creating and comparing rainfall prediction models, employing diverse machine learning algorithms, including Logistic Regression, Decision Tree Classifier, Multi-Layer Perceptron classifier (neural network), and Random Forest. The study aims not only to predict rainfall patterns but also to evaluate the performance of each model through metrics such as *Accuracy*, Cohen's kappa coefficient, and Receiver Operating Characteristic (ROC) curve analysis. Additionally, the relevance of the predictors employed in each model is thoroughly assessed. The results of extensive experimentation and analysis reveal that the Logistic Regression (*Accuracy* = 82.80 %, ROC = 82.45 %, Cohen's Kappa = 65.05 %) and Neural Network model (*Accuracy* = 82.59 %, ROC = 81.94 %, Cohen's Kappa = 64.40 %) has emerged as the most promising approach, achieving the highest percentage of accuracy, ROC and Cohen's Kappa metrics; among the models considered. This outcome underscores the effectiveness of Logistic Regression and Neural Network architectures in capturing intricate patterns and relationships within rainfall data.

1. Introduction

Climate change, characterized by heightened temperatures, regional rainfall and runoff shifts, and rising sea levels, has profound consequences. The evaluation of climate change must consider precipitation as a pivotal factor, requiring integration into hydrodynamic studies addressing its effects [1]. The intensity and frequency of rainfall are critical indicators of climate change in specific regions. On a surface level, the irregular rainfall and its variations pose threats such as floods [2,3], rising temperatures, and cyclones [4]. Predicting precipitation faces challenges due to its diverse forms, like rain, snow, and hail, which affect daily outdoor activities. The intricacies of accurate rainfall prediction are worsened by extreme climate variations [5]. The significance of rainfall forecasting has recently surged in research, given its complexities and applications in flood prediction and monitoring pollutant concentrations [6]. Factors influencing rainfall, such as humidity, temperature, and wind speed, are essential considerations [7–15]. The dependence of agriculture and water quality on daily and annual rainfall fluctuations underscores the challenges of predicting daily rainfall for effective water resource management in agriculture and water supply [16]. Consequently, weather prediction has gained popularity, leading researchers to enhance efficiency [17].

Artificial Intelligence (AI) has experienced significant growth in the digital age, leading to the emergence of applications such as Machine Learning (ML) and Deep Learning (DL) techniques widely utilized across diverse sectors [18]. In recent years, research has focused on weather prediction models, particularly the analysis of rainfall, employing various ML and DL techniques for short- and medium-term forecasts [19]. Despite the longstanding existence of rainfall forecasting, traditional methods reliant on statistical techniques have proven computationally intensive and time-consuming, offering limited efficacy [20]. Machine-learning algorithms, encompassing decision trees, K-Nearest Neighbors (KNN), linear regression, and rule-based methods, have been applied to predict rainfall, demonstrating effectiveness compared to deterministic approaches. Classical models like Logistic Regression, Decision Trees, and K-Nearest Neighbors, along with the utilization of

* Corresponding author.
E-mail addresses: mohdusmansaeedkhan@zhcet.ac.in (M. Usman Saeed Khan), 04kmsaifullah@gmail.com (K. Mohammad Saifullah), ajmal.hussain@zhcet.ac.in (A. Hussain), Hazi.Azamathulla@sta.uwi.edu, azmatheditor@gmail.com (H. Mohammad Azamathulla).

https://doi.org/10.1016/j.rineng.2024.102093
Received 6 January 2024; Received in revised form 20 March 2024; Accepted 1 April 2024
Available online 5 April 2024

Deep Learning and its subsets, have been explored as classifiers for predicting rainfall occurrences [21]. Scholars have affirmed that machine learning algorithms surpass traditional deterministic methods in weather and rainfall prediction [16].

Additionally, the application of ML extends beyond meteorology to industries like robot path planning, climatic change prediction, including cyclones and floods [22], and land cover classification [23]. Efforts have been made to enhance accuracy by assessing the correlation between rainfall and geographic coordinates alongside other atmospheric factors [24]. The learning process in rainfall prediction encompasses two types: supervised learning, involving provided features for training [25], and unsupervised learning, where learning occurs without a predefined feature set, employing neural networks for feature extraction and subsequent classification [26].

1.1. Related work

Improving the accuracy of machine learning techniques on weather forecasting has been the primary concern of many researchers over the last two decades. Some of the related studies are discussed here.

Cabezuelo [27] utilized diverse machine learning algorithms to forecast the "RainToday" variable based on the Australian rainfall dataset. The models under scrutiny encompass KNN, Decision Tree, Random Forest, and Neural Networks. In the case of KNN, the optimal settings were identified as N_neighbors: 25 and Weights: distance. Decision Tree's optimal parameters included Max_depth: 11, Max_features: 21, Class_weight: "Balanced," Criterion: "entropy" and Splitter: "best." Random Forest's most effective configuration comprised N_estimators: 30, Max_depth: 14, Max_features: 21, Criterion: "gini" and Class_weight: "balanced." Neural Networks, featuring Activation: 'relu,' Solver: 'adam, ' and Alpha: 0.05, exhibited the most favorable outcomes. Comparing these models revealed that Neural Networks achieved the highest AUC and the lowest MSE. Independently trained models enhanced accuracy when applied to specific cities (Sydney, Perth, Darwin). The inference drawn is that Neural Networks, particularly with "tanh" activation, "adam" solver, and alpha 0.05, offer the most accurate rainfall predictions in Australia. The study proposes future investigations exploring extended time frames, international datasets, and the forecasting of multiple days in advance.

Liyew and Melese [28] primarily aimed to forecast daily rainfall intensity by leveraging machine learning techniques to identify pertinent atmospheric features. The selection of environmental variables correlated with rainfall was based on Pearson correlation analysis, utilizing a threshold of 0.20. Noteworthy features such as Evaporation, Relative Humidity, Sunshine, Maximum Daily Temperature, and Minimum Daily Temperature were chosen. Multiple Linear Regression (MLR), Random Forest (RF), and XGBoost were deployed as machine learning models to predict daily rainfall quantities. Results indicated that certain attributes like year, month, day, and wind speed had negligible impact, while high correlation coefficients of 0.401 and 0.351 were observed for relative humidity and daily sunshine, respectively. The comparative analysis of model performance demonstrated that XGBoost surpassed MLR and RF, with Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) values of 3.58 and 7.85. The study concluded by endorsing XGBoost as the optimal algorithm for daily rainfall prediction.

Sethupathi et al. [29] propose a rainfall forecasting system utilizing two machine learning techniques: logistic regression and random forest in their study. The data collection for this study includes various parameters related to rainfall from different regions in India, and features such as date, location, temperature, wind, humidity, pressure, and rainfall information. The data undergoes four key preprocessing stages: data arrangement, importing dataset and libraries, removing null values, and splitting the data into training and testing sets. The machine learning algorithms, logistic regression, and random forest are then applied to the training data. The accuracy scores for logistic regression and random forest are approximately 95.9 % and 94.4 %, respectively. Both algorithms perform well, with the logistic regression algorithm being slightly more efficient. The conclusion suggests that accuracy based on these algorithms for predicting rainfall is efficient and provides accurate results, with potential for further exploration of classification methods.

Yen et al. [30] employed Echo State Network (ESN) and Deep Echo State Network (DeepESN) models for rainfall forecasting in southern Taiwan, using meteorological data from Zengwen and Yujing observatories. DeepESN outperforms ESN, exhibiting lower RMSE, Normalized Root Mean Square Error (NRMSE), and higher gamma (γ) values. The optimal training length is identified as 20,000 h. DeepESN surpasses other neural network models and maintains computational efficiency on a standard personal computer. The reduced input parameters, focusing on rainfall, air pressure, and humidity, enhance the model's performance. The study concludes that DeepESN is a reliable and efficient tool for rainfall prediction, emphasizing the significance of key parameters.

1.2. Latest studies

Latif et al. [31], assessed various predictive models for rainfall forecasting, primarily focusing on statistical models and machine learning algorithms. These models, including deep learning approaches like LSTM, utilize data from diverse sources such as satellite imagery, radar data, and ground-based observations. Overall, the study underscores the efficacy of machine learning, particularly LSTM models, in rainfall prediction across various climates and time scales, while also suggesting further exploration of remote sensing and hybrid predictive models for improved accuracy.

Ojo et al. [32] investigated the rainfall prediction to address the imperative need for planning and mitigating risks associated with variable rainfall. Two multivariate polynomial regressions (MPR) and twelve machine learning algorithms, including three artificial neural networks (ANN), four adaptive neuro-fuzzy inference system (ANFIS), and five support vector machine (SVM) algorithms, were employed to estimate monthly and annual rainfalls in a tropical locale. The proposed models utilized geoclimatic coordinates such as longitude, latitude, and altitude as input variables. Analysis based on the general performance index (c) revealed that the adaptive neuro-fuzzy inference system (ANFIS) model's algorithms outperformed the MPR, ANN, and SVM models throughout the year.

Kumar et al. [33], applied machine learning methods for forecasting rainfall in urban metropolitan cities. Time series data, characterized by temporal complexities, are utilized with a unique data segmentation approach, creating discrete training, validation, and testing sets. Two distinct models, Model-1 based on daily data and Model-2 based on weekly data, are developed and rigorously analyzed using various performance criteria. Algorithms such as CatBoost, XGBoost, Lasso, Ridge, Linear Regression, and LGBM are examined, revealing significant trends across all phases of assessment. Results indicate that ensemble-based algorithms, particularly CatBoost and XGBoost, outperform others in both models, with CatBoost demonstrating unparalleled predictive ability throughout training, validation, and testing phases.

Application of four machine learning techniques for rainfall prediction at Ranichauri station in the Tehri Garhwal district of Uttarakhand were studied [34]. The developed models were validated using statistical parameters such as root mean square error (RMSE), index of agreement (d), correlation coefficient (r), and Kling-Gupta efficiency (KGE). Results revealed that the random forest (RF) model outperformed others in both calibration and testing periods for daily and mean weekly rainfall prediction.

While machine learning (ML) models for rainfall forecasting have demonstrated potential in rainfall prediction, the complexity of rainfall formation necessitates the use of hybrid models for more accurate estimates [35]. Unlike previous studies focusing solely on individual ML models, this study includes hybrid models specifically tailored for

rainfall forecasting. These hybrid models exhibit improved accuracy and reduced uncertainty for both short and longer lead times, offering valuable insights for researchers aiming to develop precise early warning systems for rainfall forecasting.

1.3. Description of the area

The present work focuses on Aligarh city, which is situated at coordinates 27.8974° N, 78.0880° E on the western side of Utter Pradesh district which lies in Northern India, as shown in Fig. 1. According to the Indian Meteorological Department, Aligarh typically experiences rainfall during the monsoon season, spanning from late June to early October, with an average rainfall ranging between 800 mm and 900 mm. August emerges as the wettest month, presenting a 51 % chance of precipitation on a typical day.

The study is based on a comprehensive 10-year dataset from 2013 to 2022, sourced from the NASA POWER Project website, which includes various meteorological parameters relevant to rainfall prediction. These parameters encompass temperature, humidity, wind speed, precipitation, and binary indicators like RainToday and RainTomorrow. This diverse array of parameters enables a comprehensive rainfall prediction analysis using machine learning models. The dataset encompasses daily rainfall records. The primary objective is to identify the most suitable model for predicting rainfall in Aligarh. Four models, namely logistic regression, decision tree, multi-layer perceptron (MLP), and random forest, have been employed for analysis. Results indicate that the logistic regression and MLP demonstrate superior accuracy. This research contributes valuable insights into rainfall prediction models for Aligarh, aiding in the enhancement of forecasting methodologies.

In the field of meteorological studies, the region of Aligarh holds particular significance due to its distinctive climatic characteristics within the larger context of India's diverse weather patterns. The integration of geographical features, including proximity to the Himalayas and the influence of the monsoon system, contributes to a unique and complex climate in this region. This study embarks on a novel exploration of rainfall prediction in Aligarh, leveraging advanced machine learning techniques to navigate the intricacies of its weather dynamics. By exploring this new area, the research aims to unravel the complexities of precipitation forecasting, offering insights that could potentially revolutionize the understanding and prediction of rainfall patterns in the region.

2. Methodology

The methodology employed in this study encompasses a systematic and comprehensive approach to rainfall prediction using machine learning techniques, as shown in Fig. 2, which involve advanced feature selection techniques, model training, and evaluation of four distinct models.

2.1. Data processing

The primary dataset, is loaded into a Pandas Data Frame for efficient manipulation and exploration. Categorical variables, specifically "RainToday" and "RainTomorrow", were encoded numerically, with 'No' represented as 0 and 'Yes' as 1, aligning them with machine learning conventions. The 'Dates' column also undergoes a transformation into a datetime format and subsequent conversion to an 'int64' representation, enhancing temporal analysis. Additionally, a visual representation of the class distribution in the imbalanced "RainTomorrow" variable was crafted using a bar plot, where 0 denoted 'No' and 1 represented 'Yes' as shown in Fig. 3.

To detect potential missing values, a heatmap is generated using the Seaborn library as shown in Fig. 4. This visual representation offers an overview of data completeness, meaning the dataset has no missing values. If there were missing values in the dataset, then they would have been depicted with dark grey dashes in Fig. 4.

2.2. Feature selection

Advanced techniques are applied to optimize the dataset for subsequent model training. Statistical independence between individual features and the target variable, "RainTomorrow", is assessed using the chi-square method. The "SelectKBest" module retains the top five features ('Tempmin', 'Humidity', 'WindDir', 'Precipitation_MM', 'RainToday') with the highest chi-square scores, signifying their significance in predicting the target variable. Simultaneously, the random forest algorithm identifies feature importance by employing the "Gini" impurity measure. The "SelectFromModel" approach ensures the retention of features ("Tempmin", "Precipitation_MM", "RainToday") contributing

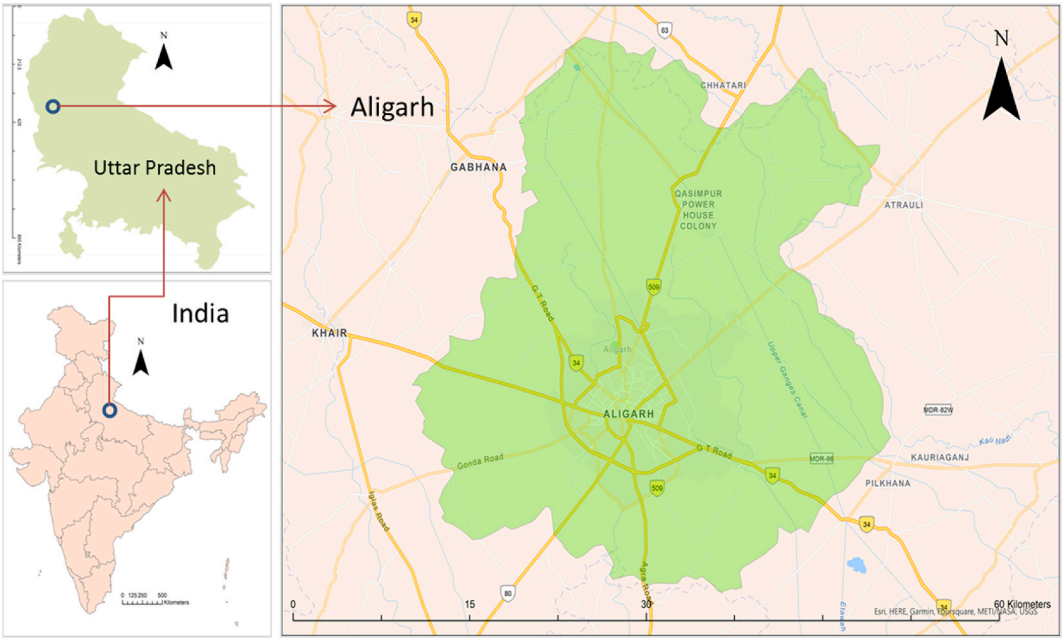


Fig. 1. Location of Aligarh city.

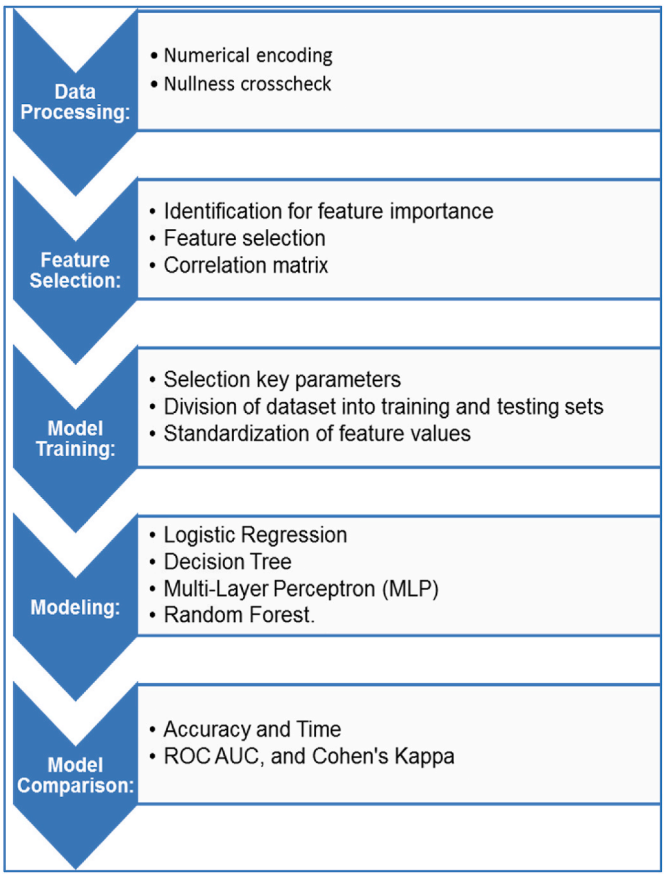


Fig. 2. Model Architecture of the present study.

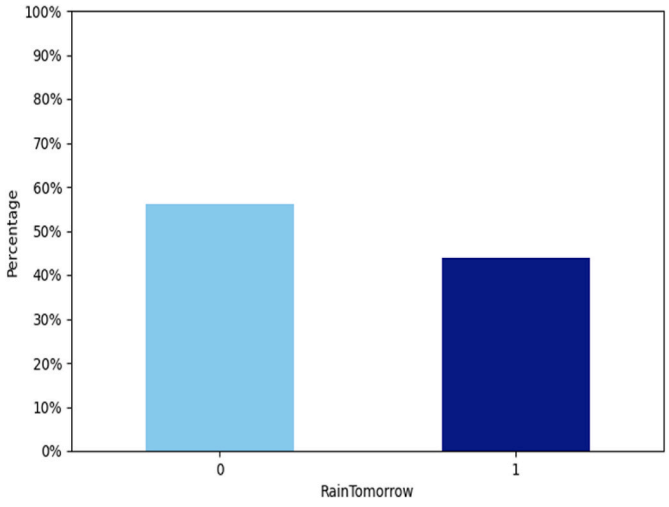


Fig. 3. Representation of rain (1) and no rain (0) in the "RainfallTomorrow" column.

significantly to model performance. This process ensures the preservation of salient attributes, enhancing the efficiency and interpretability of subsequent machine-learning models. Additionally, a correlation matrix, depicted in Fig. 5, is generated using the Seaborn library, offering a visual representation of pairwise feature correlations derived from the correlation matrix. This technical insight refines feature selection, providing an understanding of intrinsic dataset relationships and contributing to the overall robustness of subsequent analyses.

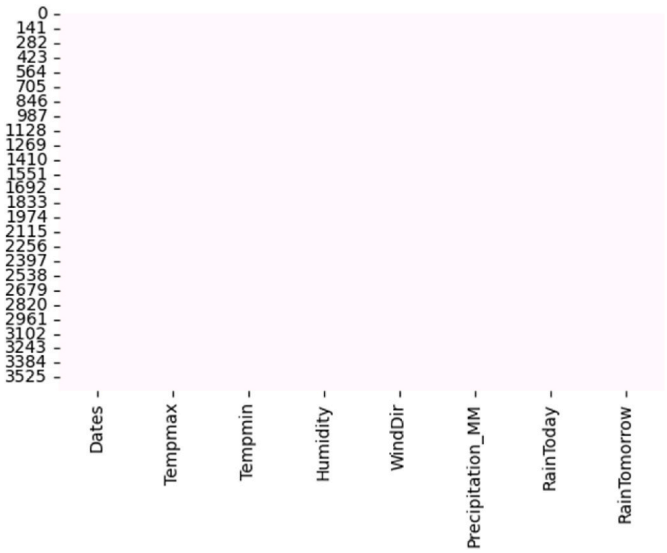


Fig. 4. Graphical representation of the data to check missing values.

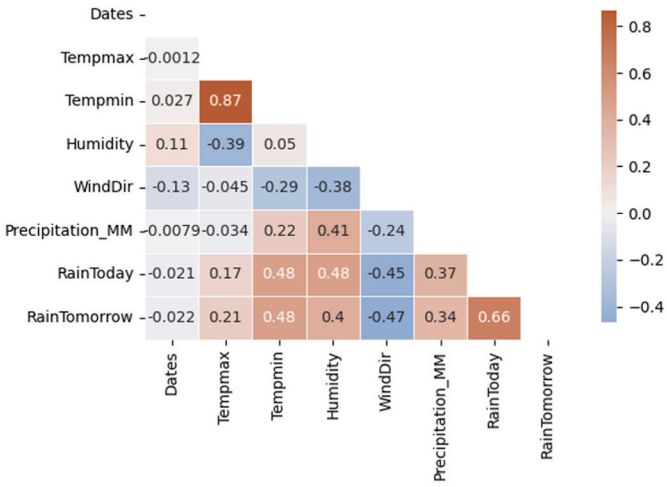


Fig. 5. Correlation matrix between features and target variables.

2.3. Model training

In the initial stages of model preparation, a strategic selection of features was undertaken, focusing on key parameters such as “Tempmin”, “Humidity”, “Precipitation_MM” and the binary variable “RainToday”. Simultaneously, the target variable, “RainTomorrow”, was identified for predictive modeling. To facilitate a robust model performance evaluation, the dataset was divided into training and testing sets using the train_test_split function from the scikit-learn library. A careful distribution was done, reserving three-fourth of the data for training (X_train and Y_train) and designating one-fourth for testing (X_test and Y_test). StandardScaler was employed to ensure uniformity and standardization of feature values to address potential variations in feature scaling. The training set features (X_train) were transformed using the fit_transform method of the scaler, and the same scaling parameters were applied to transform the testing set features (X_test). This systematic approach for feature selection, dataset partitioning, and standardized scaling was used to serve as the groundwork for subsequent model training and evaluation. It enhances overall coherence and effectiveness of the resultant machine learning framework.

2.4. Models

2.4.1. Model selection

The selection of machine learning models in this study was guided by their respective strengths and weaknesses, tailored to the specific requirements of rainfall prediction. Logistic Regression was chosen for its simplicity and interpretability, making it suitable for binary classification tasks. However, its linear decision boundary may limit its ability to capture complex relationships in the data. Decision Trees offer intuitive decision-making processes and can handle nonlinear relationships well, but they are prone to overfitting, especially with deep trees. Multi-Layer Perceptron (MLP) neural networks excel in capturing intricate patterns in data through hidden layers, yet they require careful hyperparameter tuning and are computationally expensive. Random Forests mitigate overfitting by combining multiple decision trees, offering robustness and scalability, although at the cost of decreased interpretability. Each model brings unique advantages and trade-offs, allowing for a comprehensive exploration of predictive capabilities in rainfall forecasting.

2.4.2. Logistic Regression

Logistic Regression is a generalized linear model; it uses statistical methods for binary classification problems in data science and machine learning [22]. In Python, logistic regression is implemented through a mathematical formula based on the logistic function, which transforms a linear combination of input features into probabilities. Equation (1) encapsulates the probability of an instance belonging to class 1.

$$P(Y=1)=\frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\ldots+\beta_n X_n)}} \tag{1}$$

where, *P(Y=1)* is the probability of rain (1), X_1, X_2, …, X_n are the features and β_0, β_1, …, β_n are the model coefficients. Applying logistic regression to rainfall prediction involves the preparation of historical data with features like humidity, temperature, and wind speed. The model is trained to learn the relationships between these features and the binary outcome of rainfall (1) or no rainfall (0). Consequently, the trained model predicts the rainfall probability when presented with new weather data. The model classifies instances into binary outcomes by establishing a decision threshold, making logistic regression a valuable tool for decision-making in scenarios like rainfall prediction.

2.4.3. Decision Tree

A decision tree is a supervised machine learning algorithm commonly used for both classification and regression tasks. It falls under the category of ensemble models, specifically a type of tree-based model [21]. The decision tree operates by recursively partitioning the input data into subsets based on the most significant attribute at each node. The process continues until a specified termination criterion, as shown in Fig. 6, such as a predetermined depth or a minimum number of samples in a node, is met. The decision tree framework involves nodes representing decision points, branches corresponding to the possible outcomes of decisions, and leaves representing the final predicted class or numerical value. Each decision is determined by evaluating the input features against a set of rules, typically derived from training data. The training process involves recursively optimizing the decision rules to minimize a predefined objective function, ensuring the model's ability to generalize to unseen data.

2.4.3.1. Multi-Layer Perceptron. Neural networks, specifically Multilayer Perceptron (MLP), operate through interconnected layers of nodes. Each node's output is determined by computing the weighted sum of its inputs and passing the result through an activation function. In a basic feed forward neural network with a single hidden layer, the output of a node (indexed as j) within the hidden layer can be expressed as Equation (2):

$$y_j=f\left(\sum_{i=1}^{n} w_{ij}x_i+b_j\right) \tag{2}$$

Here, y_j represents the output of node j, f is the activation function, w_{ij} denotes the weight associated with the connection between input node *i* and hidden *node j*, x_i is the input from node *i*, b_j is the bias term for node *j*, and *n* signifies the total number of input nodes. The activation function introduces non-linearity, enabling the neural network to model complex relationships in the data. This weighted summation and activation process is repeated across layers, ultimately producing the network's output. The MLP's ability to learn intricate patterns and representations makes it a powerful tool in various machine learning

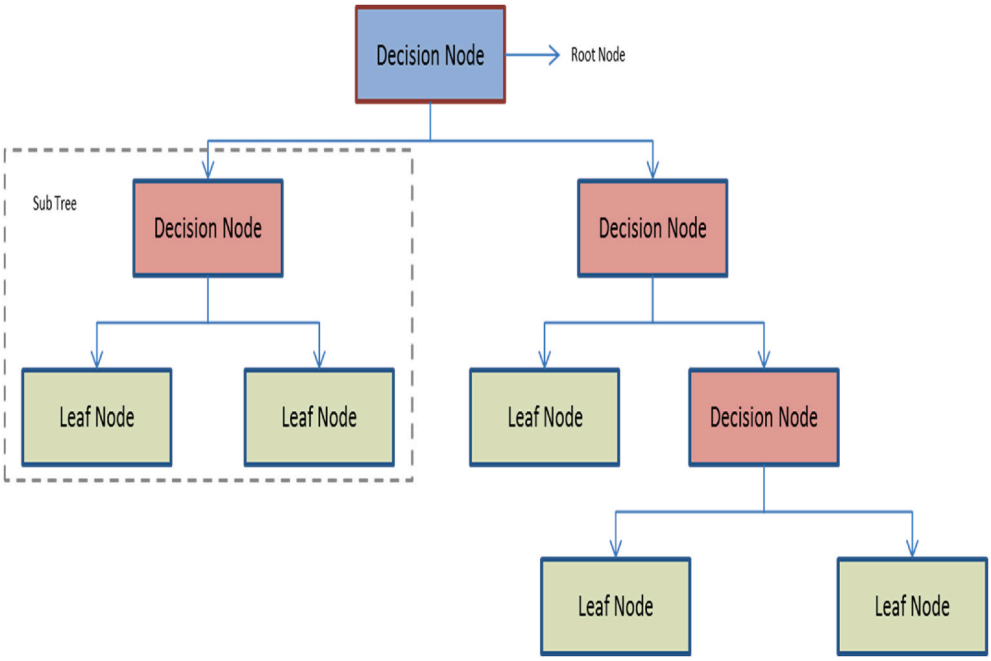


Fig. 6. Work flow of decision tree.

applications.

2.4.3.2. Random Forest. Random Forest, an ensemble learning method, constitutes a collection of decision trees, combining their predictions to enhance robustness and alleviate overfitting. Numerous decision trees are constructed during the training phase, as given in Fig. 7, each utilizing a subset of training data and features, introducing variability into the process. The resulting predictions from individual trees are aggregated through a voting or averaging mechanism [21,22]. This ensemble approach bolsters prediction accuracy and augments the model's ability to generalize to new data. The inherent diversity achieved by the random selection of data and features during tree construction contributes to the overall stability and reliability of the Random Forest, making it a widely utilized and effective tool across diverse machine learning applications.

2.5. Caliberation and validation

75 % of the data designated for training and calibration purposes was systematically selected using the train_test_split function from the scikit-learn library. This function ensures a random and representative allocation of data into the training set while maintaining the overall distribution of classes. Randomness in the selection process helps prevent biases and ensures that the training set is diverse, capturing a broad range of patterns present in the dataset. Subsequently, validation was conducted on the remaining 25 % of the data to appraise the models' performance on unseen data. Rigorous validation ensures the applicability of models beyond the training dataset, validating effectiveness of the models in real-world scenarios. Several regulatory parameters were employed to fine-tune the models for improved predictive accuracy in the calibration and validation processes. These parameters include learning rate, regularization strength, batch size, and number of iterations for optimization algorithms like gradient descent.

3. Results & discussion

Three key metrics were employed to evaluate the performance of the machine learning models in rainfall prediction. Accuracy, a fundamental metric, represents the ratio of correctly predicted instances to the total number of instances. The Logistic Regression and MLP Classifier have performed best in terms of accuracy, an in-depth analysis of precision, recall, and *f1* score was also conducted for both binary classes.

Accuracy is a measure of the overall correctness of a classification model, as given in equation (3). It calculates the ratio of correctly predicted instances (true negatives and true positives) to the total instances.

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP} \tag{3}$$

where True Positives (*TP*) are instances that are correctly predicted as positive, True Negatives (*TN*) are instances that are correctly predicted as negative, False Positives (*FP*) are instances that are incorrectly predicted as positive, and False Negatives (*FN*) are instances that are incorrectly predicted as negative.

Recall, also known as sensitivity or true positive rate, measures the ability of the model to capture all the relevant instances. It is the ratio of true positives to the total number of actual positive instances as shown in equation (4).

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

Precision, also known as positive predictive value, measures the accuracy of the positive predictions; its mathematical representation is shown in equation (5). It is the ratio of true positives to the total number of instances predicted as positive.

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

The *f*1 score is the harmonic mean of precision and recall. It provides a balance between precision and recall, and it is especially useful when there is an uneven class distribution, as given in equation (6).

$$f1\ score = \frac{2 * (precision * recall)}{precision + recall} \tag{6}$$

The logistic regression model demonstrated commendable

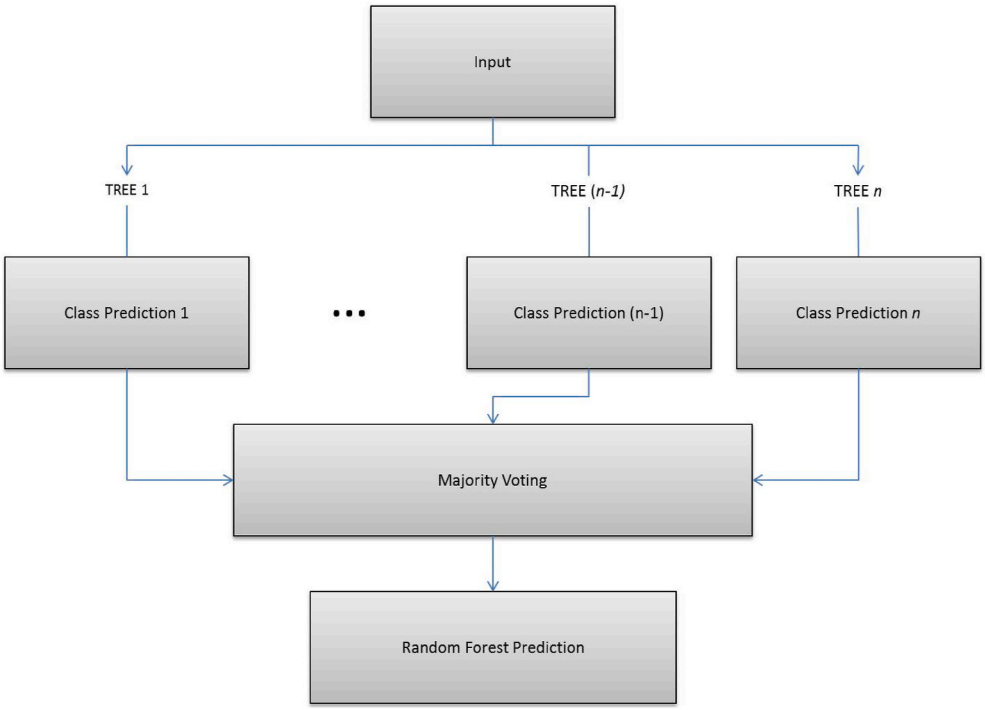


Fig. 7. Work flow of random forest.

performance in predicting instances of positive rainfall, with a *Precision* of 81.22 %, a *Recall* of 79.40 %, and an *f*1 score of 80.30 %. Similarly, the decision tree model exhibited competitive metrics for positive instances, showcasing a *Precision* of 80.77 %, a *Recall* of 67.74 %, and an *f*1 score of 73.68 %. The multi-layer perceptron (MLP) model demonstrated noteworthy precision, recall, and *f1* score values for positive rainfall instances, with scores of 82.80 %, 79.40 %, and 80.30 %, respectively. Lastly, the random forest model yielded substantial performance in predicting positive rainfall, with a *Precision* of 82.58 %, a *Recall* of 72.95 %, and an *f*1 score of 77.47 %. These metrics collectively provide an understanding of the models' abilities to identify positive instances accurately. The highest value among the *Precision*, *Recall*, and *f1* scores for positive rainfall instances is achieved by the MLP model. The lowest value among these metrics is seen in the decision tree model. Based on these values, the MLP model appears to be the optimum model for predicting positive rainfall instances, as it achieves the highest performance across all three evaluation metrics.

Receiver Operating Characteristic (ROC) Area under Curve (AUC), a validation curve, assesses the model's ability to discriminate between positive and negative instances. It can be seen in Fig. 8 that the orange line typically represents the ROC curve for a specific model or classifier, illustrating the trade-off between the true positive rate (sensitivity) and the false positive rate. The blue dotted line, on the other hand, represents the ROC curve for a random or baseline model. The orange line deviates significantly above the diagonal (towards the top-left corner), indicating better-than-random performance in each case, with the degree of deviation provides insights into the model's effectiveness in distinguishing between the classes. The area under the ROC curve (AUC) quantifies this overall discriminative power, with the largest AUC in (a), which is Logistic Regression's performance and the smallest AUC in the case of (b), which has been generated for Decision Tree. Also, (c) and (d) show the plots of MLP and Random Forest, respectively.

Cohen's Kappa, a statistic that considers the agreement between predicted and actual outcomes while accounting for chance, measures the model's performance beyond what would be expected by random chance. It has been found that Logistic Regression and MLP Classifier have performed best in terms of Cohen Kappa, as shown in Table 1.

The comparison (validation) between all four machine learning models Logistic Regression, Decision Tree, Multi-Layer Perceptron (MLP), and Random Forest, is visually depicted through two graphs, providing an insightful representation of their performance. Fig. 9 presents a comparative analysis of accuracy to assess the overall precision of predictions. Here, the red line shows the accuracy of different models, and the bars depict time. Overall, Logistic Regression and MLP exhibit slightly superior accuracy.

Fig. 10, a validation presentation, illustrates the ROC AUC values in the form of bars for each model; it delves into the models' discriminative capabilities. Additionally, the red line in the graph corresponds to Cohen's Kappa scores, offering a distinct perspective on the models' performance. Logistic Regression and MLP were found out to be of a

Table 1
Comparison of the key Metrics.

Metrics	Logistic Regression	Decision Tree	MLP	Random Forest
Accuracy	82.80 %	78.64 %	82.59 %	81.27 %
ROC_AUC	82.45 %	77.50 %	81.94 %	80.40 %
Cohen's Kappa	65.05 %	55.94 %	64.40 %	61.55 %.

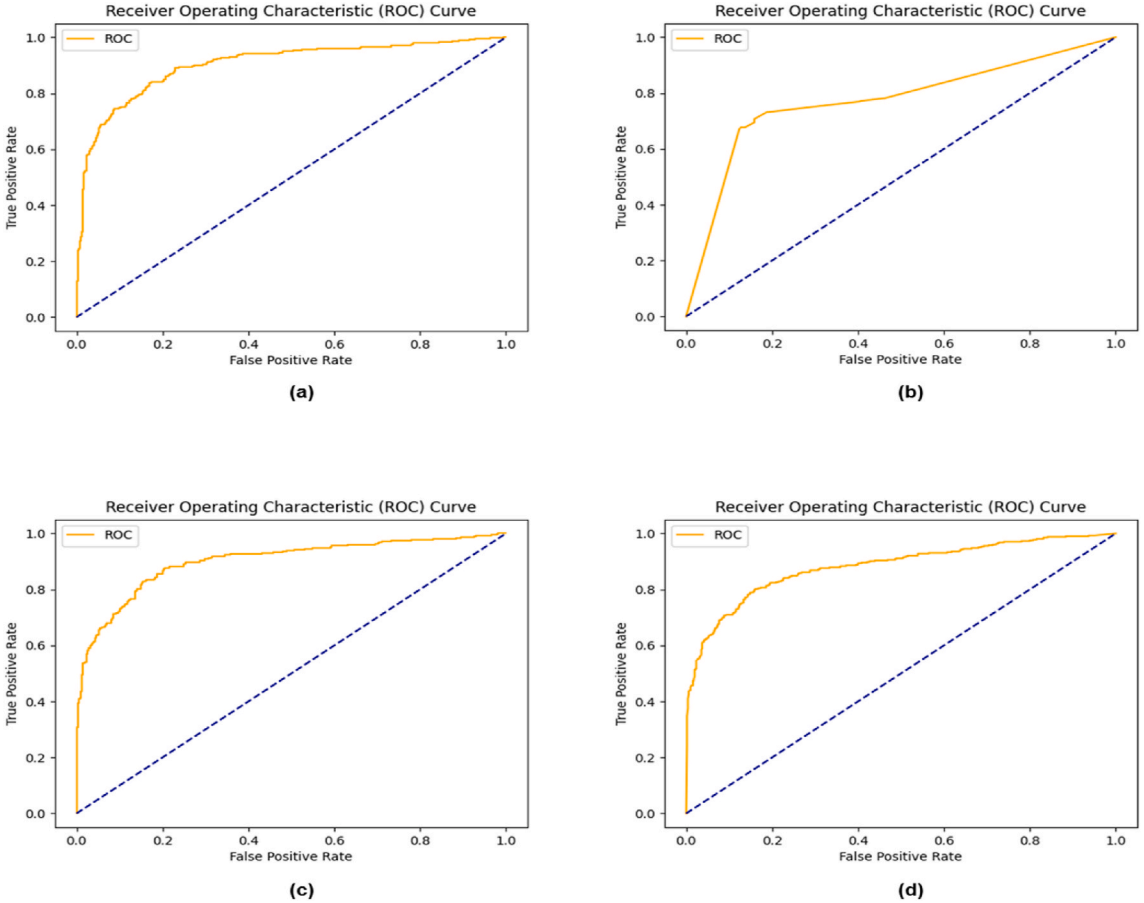


Fig. 8. Receiver operating curves.

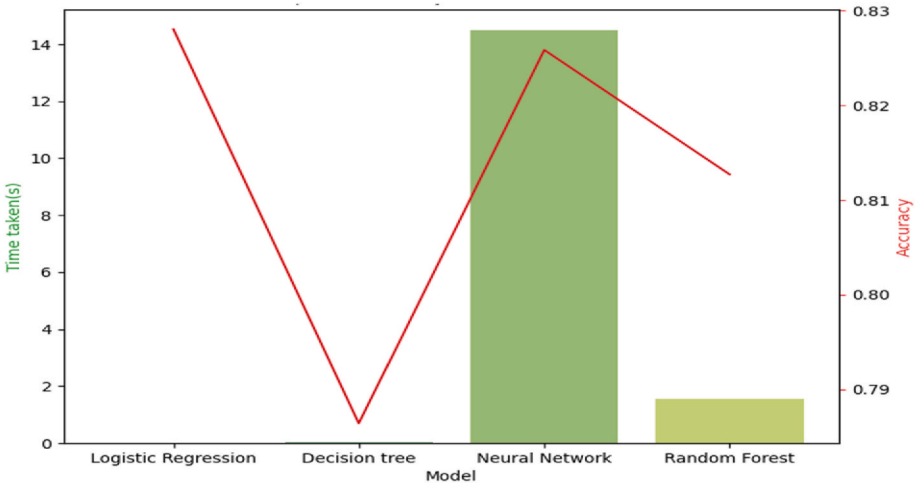


Fig. 9. Model comparison for accuracy.

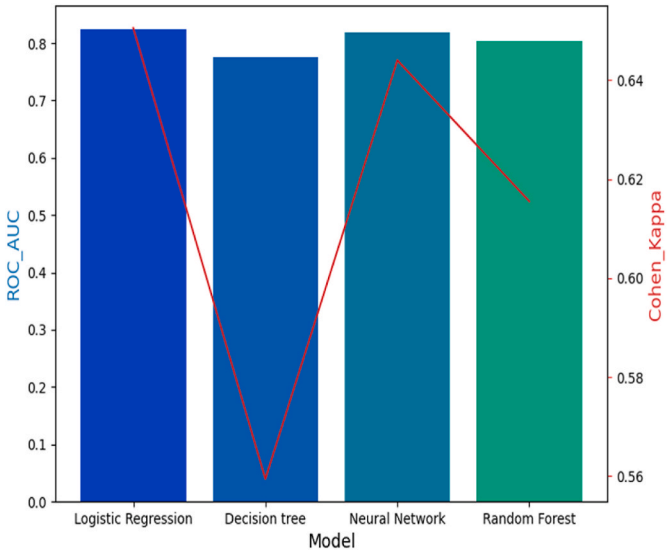


Fig. 10. Model comparison for ROC and Cohen's kappa.

higher true positive rate for various false positive rates. This visually emphasizes their superior ability to distinguish between positive and negative instances, a critical factor in accurate rainfall prediction scenarios. Together, these metrics offer a comprehensive evaluation of the models' accuracy, discriminatory power, and agreement with actual outcomes in the context of rainfall prediction. The study's results give detailed insights into how to predict rainfall using Machine Learning models. It helps us understand the complicated workings of using advanced computer programs to forecast the weather accurately. This understanding is important for making weather predictions more reliable and precise.

4. Conclusion

The study delves into the intricate dynamics of rainfall prediction through meticulous examination of machine learning models, including Logistic Regression, Decision Tree, Multi-Layer Perceptron (MLP), and Random Forest. By leveraging advanced techniques in data processing, feature selection, and model training, the research endeavors to enhance the accuracy and reliability of weather forecasting methodologies. The comprehensive evaluation of model performance, elucidated through quantitative metrics such as accuracy, precision, recall, and F1 score, underscores the effectiveness of the proposed approaches in capturing complex patterns inherent in meteorological data. Moreover, the utilization of macro and weighted averages provides nuanced insights into model efficacy, particularly in addressing class imbalances prevalent in rainfall prediction scenarios. Ultimately, the overarching goal of the article is to contribute to the advancement of predictive modeling techniques in weather forecasting, thereby enabling more reliable and precise predictions essential for informed decision-making in various domains. Through the integration of statistical analysis, model evaluation, and a clear delineation of research objectives, the conclusion encapsulates the essence of the study and its implications for the broader scientific community.

CRediT authorship contribution statement

Mohd Usman Saeed Khan: Formal analysis, Data curation, Conceptualization. **Khan Mohammad Saifullah:** Writing – original draft, Software, Data curation, Conceptualization. **Ajmal Hussain:** Conceptualization, Data curation, Supervision, Writing – review & editing. **Hazi Mohammad Azamathulla:** Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] T.A. Duong, M.D. Bui, P. Rutschmann, A comparative study of three different models to predict monthly rainfall in Ca Mau, Vietnam, in: Wasserbau-Symposium Graz, 2018.
[2] A.G. Yilmaz, The effects of climate change on historical and future extreme rainfall in Antalya, Turkey, Hydrol. Sci. J. 60 (12) (2015) 2148–2162, https://doi.org/10.1080/02626667.2014.945455.
[3] M.U.S. Khan, M. Abdullah, A.A. Khan, *Flood Modelling and Simulation using HEC-RAS* (No. EGU23-5564), Copernicus Meetings (2023), https://doi.org/10.5194/egusphere-egu23-5564.
[4] Y.Y. Loo, L. Billa, A. Singh, Effect of climate change on seasonal monsoon in Asia and its impact on the variability of monsoon rainfall in Southeast Asia, Geosci. Front. 6 (6) (2015) 817–823, https://doi.org/10.1016/j.gsf.2014.02.009.
[5] A.U. Rahman, S. Abbas, M. Gollapalli, R. Ahmed, S. Aftab, M. Ahmad, A. Mosavi, Rainfall prediction system using machine learning fusion for smart cities, Sensors 22 (9) (2022) 3504, https://doi.org/10.3390/s22093504.

[6] A.Y. Barrera-Animas, L.O. Oyedele, M. Bilal, T.D. Akinosho, J.M.D. Delgado, L. A. Akanbi, Rainfall prediction: a comparative analysis of modern machine learning algorithms for time-series forecasting, Machine Learning with Applications 7 (2022) 100204, https://doi.org/10.1016/j.mlwa.2021.100204.
[7] H.A. Elwell, M.A. Stocking, Rainfall parameters and a cover model to predict runoff and soil loss from grazing trials in the Rhodesian sandveld, Proceedings of the Annual Congresses of the Grassland Society of Southern Africa 9 (1) (1974) 157–164, https://doi.org/10.1080/00725560.1974.9648736.
[8] H.M. Azamathulla, U. Rathnayake, A. Shatnawi, Gene expression programming and artificial neural network to estimate atmospheric temperature in Tabuk, Saudi Arabia, Appl. Water Sci. 8 (2018) 184, https://doi.org/10.1007/s13201-018-0831-6.
[9] A. Perera, U. Rathnayake, Impact of climate variability on hydropower generation in an un-gauged catchment: Erathna run-of-the-river hydropower plant, Sri Lanka, Appl. Water Sci. 9 (2019) 57, https://doi.org/10.1007/s13201-019-0925-9.
[10] B. Khaniya, C. Karunanayake, M.B. Gunathilake, U. Rathnayake, Projection of future hydropower generation in Samanalawewa power plant, Sri Lanka, Math. Probl Eng. 2020 (2020) 8862067, https://doi.org/10.1155/2020/8862067. , 11 pages.
[11] C. Karunanayake, M.B. Gunathilake, U. Rathnayake, Inflow forecast of Iranamadu Reservoir, Sri Lanka, under Projected climate scenarios using artificial neural networks, Applied Computational Intelligence and Soft Computing 2020 (2020) 8821627, https://doi.org/10.1155/2020/8821627. , 11 pages.
[12] L. Mampitiya, N. Rathnayake, Y. Hoshino, U. Rathnayake, Performance of machine learning models to forecast PM10 levels, MethodsX (2024) 102557, https://doi.org/10.1016/j.mex.2024.102557.
[13] S.S. Balkisson, S.R. Gunakala, H.M. Azamathulla, Areal precipitation and depth-area duration curves for regions in trinidad using a triangulated grid, ISSN 1112-3680, Larhyss J. (56) (2023). Dec 2023, pp. 235–265.
[14] T. Jayathilake, R. Sarukkalige, Y. Hoshino, U. Rathnayake, Wetland water level prediction using artificial neural networks—a case study in the colombo flood detention area, Sri Lanka, Climate 11 (1) (2022) 1.
[15] A. Danladi, M. Stephen, B.M. Aliyu, G.K. Gaya, N.W. Silikwa, Y. Machael, Assessing the influence of weather parameters on rainfall to forecast river discharge based on short-term, Alex. Eng. J. 57 (2) (2018) 1157–1162, https://doi.org/10.1016/j.aej.2017.03.004.
[16] K. Namitha, A. Jayapriya, G.S. Kumar, Rainfall prediction using artificial neural network on map-reduce framework, in: Proceedings of the Third International Symposium on Women in Computing and Informatics, 2015, pp. 492–495, https://doi.org/10.1145/2791405.2791468.
[17] S.E. Haupt, J. Cowie, S. Linden, T. McCandless, B. Kosovic, S. Alessandrini, Machine learning for applied weather prediction, in: 2018 IEEE 14th International Conference on E-Science (E-Science), IEEE, 2018, pp. 276–277.
[18] P. Asha, A. Jesudoss, S.P. Mary, K.S. Sandeep, K.H. Vardhan, An efficient hybrid machine learning classifier for rainfall prediction, in: Journal of Physics: Conference Series, IOP Publishing, 2021 012012 (Vol. 1770, No. 1.
[19] Q.B. Pham, T.C. Yang, C.M. Kuo, H.W. Tseng, P.S. Yu, Combing random forest and least square support vector regression for improving extreme rainfall downscaling, Water 11 (3) (2019) 451, https://doi.org/10.3390/w11030451.
[20] P. Singh, B. Borah, Indian summer monsoon rainfall prediction using artificial neural network, Stoch. Environ. Res. Risk Assess. 27 (2013) 1585–1599, https://doi.org/10.1007/s00477-013-0695-0.
[21] S. Ghosh, M.K. Gourisaria, B. Sahoo, H. Das, A pragmatic ensemble learning approach for rainfall prediction, Discover Internet of Things 3 (1) (2023) 13, https://doi.org/10.1007/s43926-023-00044-3.
[22] G. Ireland, M. Volpi, G.P. Petropoulos, Examining the capability of supervised machine learning classifiers in extracting flooded areas from Landsat TM imagery: a case study from a Mediterranean flood, Rem. Sens. 7 (3) (2015) 3372–3399, https://doi.org/10.3390/rs70303372.
[23] C. Huang, L.S. Davis, J.R.G. Townshend, An assessment of support vector machines for land cover classification, Int. J. Rem. Sens. 23 (4) (2002) 725–749, https://doi.org/10.1080/01431160110040323.
[24] C.L. Wu, K.W. Chau, Prediction of rainfall time series using modular soft computingmethods, Eng. Appl. Artif. Intell. 26 (3) (2013) 997–1007, https://doi.org/10.1016/j.engappai.2012.05.023.
[25] S.B. Kotsiantis, I. Zaharakis, P. Pintelas, Supervised machine learning: a review of classification techniques, Emerging artificial intelligence applications in, computer engineering 160 (1) (2007) 3–24.
[26] S. Divya, P. Asha, Earlier diagnosis and survey of diabetes mellitus using machine learning techniques, in: 2019 Third International Conference on Inventive Systems and Control (ICISC), IEEE, 2019, pp. 37–38.
[27] A. Sarasa-Cabezuelo, Prediction of rainfall in Australia using machine learning, Information 13 (4) (2022) 163, https://doi.org/10.3390/info13040163.
[28] C.M. Liyew, H.A. Melese, Machine learning techniques to predict daily rainfall amount, Journal of Big Data 8 (2021) 1–11, https://doi.org/10.1186/s40537-021-00545-4.
[29] M.G. Sethupathi, Y.S. Ganesh, M.M. Ali, Efficient rainfall prediction and analysis using machine learning techniques, Turkish Journal of Computer and Mathematics Education (TURCOMAT) 12 (6) (2021) 3467–3474, https://doi.org/10.17762/turcomat.v12i6.7135.
[30] M.H. Yen, D.W. Liu, Y.C. Hsin, C.E. Lin, C.C. Chen, Application of the deep learning for the prediction of rainfall in Southern Taiwan, Sci. Rep. 9 (1) (2019) 12774, https://doi.org/10.1038/s41598-019-49242-6.
[31] S.D. Latif, N.A.B. Hazrin, C.H. Koo, J.L. Ng, B. Chaplot, Y.F. Huang, A.N. Ahmed, Assessing rainfall prediction models: exploring the advantages of machine learning and remote sensing approaches, Alex. Eng. J. 82 (2023) 16–25.
[32] O.S. Ojo, S.T. Ogunjo, Machine learning models for prediction of rainfall over Nigeria, Scientific African 16 (2022) e01246.
[33] V. Kumar, N. Kedam, K.V. Sharma, K.M. Khedher, A.E. Alluqmani, A comparison of machine learning models for predicting rainfall in urban metropolitan Cities, Sustainability 15 (18) (2023) 13724.
[34] S. Markuna, P. Kumar, R. Ali, D.K. Vishwkarma, K.S. Kushwaha, R. Kumar, A. Kuriqi, Application of innovative machine learning techniques for long-term rainfall prediction, Pure Appl. Geophys. 180 (1) (2023) 335–363.
[35] S.Q. Dotse, I. Larbi, A.M. Limantol, L.C. De Silva, A review of the application of hybrid machine learning models to improve rainfall prediction, Modeling Earth Systems and Environment 1–26 (2023).