Mehran University Research Journal of Engineering and Technology

https://doi.org/10.22581/muet1982.2232

2024, 43(3) 1-14

Comparative evaluation of machine learning algorithms for rainfall prediction to improve rice crops production

Beenish Ayesha Akram [a,*], Amna Zafar [b], Talha Waheed [b], Khaldoon Khurshid [b] , Tayyeb Mahmoood [c]

[a] *Department of Computer Engineering, University of Engineering and Technology, Lahore*

[b] *Department of Computer Science, University of Engineering and Technology, Lahore*

[c] *Department of Electrical Engineering, Rachna College of Engineering and Technology, Lahore*

[*] Corresponding Author: Beenish Ayesha Akram, Email: beenish.ayesha.akram@uet.edu.pk

Received: 25 August 2021, Accepted: 03 June 2024, Published: 01 July 2024

KEYWORDS

Rainfall Prediction

Machine Learning

Agriculture

Climate Effects on Crops

Machine Learning in Agriculture

ABSTRACT

Rainfall has a huge impact on agriculture because it's one of the key causes of crops devastation. Farmers face a slew of issues when unexpected heavy rains fall, as their planted crops are washed away or damaged. Pakistan is an agricultural country where new methods and techniques are needed to improve the traditional farming methods. This research intends to provide aid in the protection of crops from severe rains using machine learning to accurately anticipate the possibility of rainfall, which is a well-known agricultural problem. Various weather factors such as temperature, humidity, and atmospheric pressure can be used to predict rainfall patterns. Rainfall prediction can be used to identify and furnish future rainfall descriptions for agricultural planning for food security, allowing farmers to take precautionary measures to safeguard rice fields. Naïve Bayes, LogitBoost, RIPPER, Decision Stump, AdaBoost, Random Forest, Artificial Neural network, and K* were evaluated for rainfall prediction based on accuracy, precision, recall, F1-measure, Root Mean Squared Error, area under receiver operating characteristic curve, elapsed training time and elapsed testing time. The results obtained indicate that the best performance is achieved by Random forest with maximum accuracy of 83.2%, followed by ANN (82.5%), LogitBoost (82.2%), RIPPER (82%), naïve Bayes (80.3%), AdaBoost (80.2%), and K*(79.2%) respectively. K* The lazy approach involved a minimum of training time and a maximum of test time. Maximum training time was consumed by Random Forest and minimum testing time was taken by Decision Stump.

1. Introduction

Agriculture was the key progress among human evolution that provided means for feeding the growing population in urban regions. Agricultural techniques and methods evolution chronicles back to ancient civilizations. It is estimated that the world population will grow to 9.2 billion people by the year 2050 [1]. This rapid rise in population demands global food production to be increased by 70%. Hence, it is essential to improve agriculture management practices. Modern horticulture techniques incorporating computer aided plantation and crop monitoring can be helpful in better crop yields. Farmers employing advanced agricultural methods can prevent crop damage more efficiently than using simply traditional methods. The crops production heavily depends upon variables like precipitation and temperature.

Direct-seeded fields get influenced by overwhelming precipitation like rice crops have destitute plant stand. Crops contrast in their resilience to waterlogging. Rice areas can tolerate rain to some extent but if the level of precipitation goes past the limit, it severely influences the field. Rice is primarily grown in bolstered zones that achieve heavy annual rainfall of some limit. It demands a temperature of around 25°C and rainfall of almost close to 115 cm. As well much rain and overcast skies can moderate down the process of growth and influence bloom or inevitably cause devastation of crops. If 125 mm of rain is gotten in two consecutive hours, it is called overwhelming precipitation. Rainfall at the stem prolongation stages emphatically impacts the rice plants, in the long run, increment the rate of plantation, increments rice production. But its negative impact is observed at heading and blooming stages. Therefore, farmers should take precautionary measures to avoid the rain that could damage rice crops when precipitation exceeds the cap.

Rainfall foretelling is an extremely vital element of serious and irregular rain that will have several impacts like the destruction of rice crops, harm to property and farms. Thus, improved rain prediction models are vital for generation of early warning that can minimize the risk of damaging crops and conjointly managing the agricultural croft efficiently. The prediction mainly helps agronomists and also water resources are efficiently utilized. The rain prediction may be a strenuous task and therefore the results ought to be faultless. Different hardware devices have been used for envisioning rainfall by utilizing the atmospheric conditions such as temperature, wind, sunshine, humidity, and pressure. Based on such sensory inputs from the crop fields machine learning techniques can be employed to accurately predict rainfall. Moreover, a historical analysis of rain can also be used to predict the rain for a future season. Our work is focused on finding the most suitable machine learning approach for rainfall prediction. Section 2 discusses related work. Methodology is presented in section 3, and results are discussed in section 4. Finally, conclusion and future work are summarized in section 5.

1.1 Contribution

a) This research is focused on developing and implementing a machine learning-based rainfall prediction system to enhance agricultural management. It began by compiling and cleaning a comprehensive dataset of key weather variables, ensuring data integrity. Relevant features were utilized to capture significant patterns, aiding in accurate predictions.
b) Various machine learning models were evaluated, including linear regression, decision trees, random forests, SVM, neural networks, and ensemble methods, selecting the most robust model. These models were trained and fine-tuned using cross-validation and rigorously assessed their performance with metrics like MAE, RMSE, and R^2 to ensure accuracy.
c) A comprehensive literature review was conducted for historical analysis of rainfall patterns, guiding model development and assessing the impact of accurate rainfall prediction on agricultural productivity. Future refinements are suggested in the future work section, including incorporating more data sources, improving feature engineering, and scaling the system to predict other critical weather variables.

2. Related Work

The scope of the research laid out in this paper concerns the use of several machine learning (ML) algorithms for prediction of the rainfall in the rice crop fields so that the farmers can take precautionary measures beforehand in order to save the crops from the devastation of the rain. Following are some work published recently in the field of agricultural practices using machine learning approaches. The research in [2] used Remote Piloted Aircraft Systems (RPAS) to develop predictive models for estimating indirect nitrogen levels and grain yield in irrigated rice. The models used multispectral images, 11 vegetation indices, Spearman's correlation coefficient, and the Multi-Layer Perceptron algorithm for performance evaluation. The Spearman correlation shows optimal rice monitoring window for RPAS occurs during reproduction phase. Machine Learning (MLP) generates more accurate models for Narea, with good MLP at all stages and excellent accuracy. This combination is efficient for precision agriculture in irrigated rice fields.

M. Mohammed et al. [3] discussed that the unpredictable precipitation causes incredible devastation of crops and ranches. To foresee rainfall, they apply Regression analysis with dependent and independent variables. They proposed its usage for testing the relationship between dependent and independent variables for rainfall prediction, they apply Support Vector Regression, Lasso Regression, and Multiple Linear Regression. They concluded that Lasso Regression has the most elevated mean absolute error than the other two models. They concluded that tuned SVR model that provided the best results.

R. Aguasca-Colomo et al. [4] developed and compared several precipitation prediction models already established using machine learning techniques for the island of Tenerife. They applied several data mining algorithms, hybrid global climate model, wind gust prediction, and numerical weather prediction models. They concluded that XGBoost (Extreme Gradient Boosting) showed the best performance.

U. Shah et al. [5] applied several ML algorithms for the prediction of rainfall and expound strategies utilized. They showed that Auto Regressive Integrated Moving Average (ARIMA) method had Neural Network for minimum temperature, wind speed and the minimum Root Mean Squared Error (RSME) for maximum temperature. For forecasting relative humidity, Support Vector Regression had the lowest RSME. Whereas, Random Forest had the most noteworthy accuracy of 70.5%.

M Rokonuzzaman et al. [6] reported results on their local regions suitability for crops. They discussed that Rangpur region was found to be best favourable for rice generation. They collected month wise and yearly precipitation data from various divisions spanning years 1983-2013. The research was primarily focused on precipitation mentioning it as the foremost prevailing element for rice production. Precipitation and rice production are emphatically interrelated. They identified that rice production was maximized provided precipitation was in the appropriate range i.e.800-2500 mm, but less than typical precipitation (<1800 mm) diminished the rice production. Whereas, when the precipitation was at the most elevated level (i.e. 2500<) the overall entire rice production was decreased overall in the cultivated regions due to overwhelming precipitation coming about in surges.

S. Sujariya et al. [7] analysed the rainfall variability in Northeast Thailand for the duration of year 2000–2015. They aimed to find the rainfall patterns and how they affected the length of the rice-growing time period and crop yield by utilizing a machine learning model. No substantial change was discovered for yearly, mid and late-season rainfall, but a substantial decrease was discovered in the quantity of early season rainfall throughout the 16 years in all groupings. The research established that the beginning and the final stage of the rice growing time period were delayed with decreased early rainfall over the period of studied 16 years whereas they did not observe any significant change in the total duration of the rice growing time period. They advised that the adjustment to varying rainfall patterns demands for pre-planning for yield maximization.

S. M. K. Hassan [8], reported their results on the effect of flooding on rice production. Their experimental area was in Bangladesh. They utilized two variants of an econometric model, i.e. a total production model, and a yield model. The first model showed great deal of production. On the other hand, the yield model employed the log of yield in terms of tons per acre. They also reported results generated by the assembly model indicated the vulnerability of the boro variant of rice crops incorporating meaningful coefficients along with flood damage indicator variables. The vulnerability spatial dimensions became evident as there were some regions/districts which had more detrimental effects on various kinds of rice although the national level approximations did not reveal the fact. Due to multicollinearity, the yield model employed alike pre-normalized variables to the production model. The research suggested that although, the later had theoretically appealing attributes. However, its results were not substantial. The apparent reason for that was merely one the features/variables produced the expected effects.

N. Oswal et al. [9] demonstrated various experiments that required the usage of several ML techniques to generate models for rainfall prediction. They covered major cities of Australia obtaining their data for the study. This study was split up into three categories 1) modelling inputs, 2) modelling methods, and 3) pre-processing techniques. They applied models such as Logistic Regression (LR), k-Nearest Neighbours (k-NN), Rule-based learning, Decision Tree (DT) and Ensembles to predict rainfall. They presented results demonstrating the comparison of these ML techniques and their suitability for rainfall prediction. No particular model was ruled out as the winner for rainfall prediction because the classifiers' performance varied heavily with changing input datasets.

M. Yen et al. [10] proposed using a deep neural network for rainfall prediction. They trained a prediction model using different potential predictors/features to estimate the rainfall in southern Taiwan. They forecasted rainfall by employing the Deep Echo State Network (ESN/DeepESN) model. The meteorological dataset was taken from the two observational stations and the Sea Level Centre. They considered temperature, atmospheric pressure, precipitation, wind speed, humidity, wind direction, and water level, a total of seven features for the prediction. They evaluated the impact of each and every input feature by taking an input feature off on rotation supported by the DeepESN model. Results demonstrated that the DeepESN proved to be better model to predict rainfall in comparison with the other models.

Rahman et al. [11] described that floods are a significant issue in Bangladesh, impacting agricultural production and livelihood well-being. Rice, the country's most important crop, is affected by flooding. They found that a 22% flood threshold is necessary for rice area coverage and production. Up to this threshold, a one-square-kilometer increase in flooding would increase rice area coverage by 31 hectares, while production would increase by 492 tons. However, a one-square-kilometer increase above this threshold would reduce rice production by 70 tons. The study suggests government support, stress coping strategies, and the development of stress-tolerant, high-yielding rice varieties.

Y. Chen et al. [12] proposed employing high-resolution climate model for the summer rainfall prediction in the United Kingdom. During summer season in UK, twelve-member convection-permitting ensemble (CPM) ensemble was employed to evaluate the global climate change effect on hourly rain. Physical characteristics were obtained to predict future trends and then past and future prediction periods were compared. The CPM simulation for the duration of years ranging from 1980 to 2000 and for future years 2060–2080 were considered. The model was validated using various datasets, which included Radar data and hourly rainfall approximations. The authors drew comparisons between the past period and the future period over three geographical regions across the country (North West: NW, North East: NE and South: S). Each of these three regions had their own distinct climatology. The outcome predicted for drier and lesser rainy future summers but with more intense rainfalls.

E. Putri [13] evaluated various ML methods for Rainfall forecasting and prediction. Their main focus was on climate change detection along with finding the pattern of its relationship to rice production. They proposed employing Extreme Value Theory (EVT) to estimate the behaviour of the rainfall dispersion because the rainfall dataset contained extreme values and heavy tail data distribution. EVT is extensively utilized in various fields, e.g. finance, climate change, risk management, engineering etc.

M. Bagbohounan [14] devised and compared several prediction models for the temperature estimations in the lower stream Region of Republic of The Gambia. The economy of Republic of The Gambia heavily relies on rain-fed agriculture and related services. Temperature changes can be detrimental for economic stability within the country. Temperature and downfall values were analysed for the period of 1943 to 1983 which showed a rise of 1°C in the annual mean temperature and two hundredth to twenty fifth decrease in the annual average rainfall. Rice is the main basic food of the country (60-70% consumption). Hence drastic decrease in rice production can heavily impact the food security degree of rural homes and even the urban areas.

Y. Guo et al. [15] utilized various strategies for rice yield prediction in East China by employing Artificial Neural Networks and Partial method of least squares Regression. Crop growth models used bio-physical features that contain genotypes, weather, soil conditions, and crop management approaches, are employed conjointly to produce strategies for crop yield management. Several different information sources such as remote sensing information, climate information, soil data, and science traits, are used for prediction of the rice yield. Moreover, experimental data obtained from remote systems and environmental information from native climate stations are usually incorporated to prediction rainfall.

L. Wickramasinghe et al. [16] several statistical techniques along with machine learning algorithms for identification of the relationship among past climate variables and crop yield. They utilized regression techniques, artificial neural networks (ANNs), support vector machines (SVMs) to establish the association between climate variables and resulting rice yield. Climate parameters such as minimum temperature, precipitation, mean temperature, maximum temperature, and related crop evapotranspiration along with past four-year yield history were utilized with ML algorithms in order to predict future crop yields. They concluded that a hybrid MLR-ANN model provided maximum accuracy in comparison with the traditional models such as MLR, random forest (RF), ANN, support vector regression (SVR), k-Nearest Neighbour (KNN), and SVM.

J. Jambo [17] conducted a statistical analysis of rainfall variability and its impact on wheat production in the Agarfa District of Ethiopia. They employed regression techniques to assess the relationship between rainfall patterns and wheat yield, finding significant coping strategies that farmers could adopt to mitigate the adverse effects of rainfall variability. The study highlighted the importance of pre-planning and adaptive measures for maintaining crop productivity.

M. Müller et al. [18] investigated the morphological and physiological responses of Calobota sericea plants under conditions of water limitation and subsequent sprinkling. They applied a combination of statistical methods and machine learning algorithms to analyse the plant responses. The study concluded that specific irrigation strategies significantly improved the resilience of these plants to drought conditions, suggesting potential applications in drought-prone agricultural regions.

A. Getahun et al. [19] performed trend and change-point detection analyses of rainfall and temperature over the Awash River basin in Ethiopia. They utilized advanced time series models to identify significant changes in climate variables and their impacts on local agriculture. Their findings underscored the need for adaptive agricultural practices in response to changing climatic conditions.

A. Jayachandran [20] studied wage responses to productivity shocks in developing countries, focusing on labour markets within agricultural settings. They applied econometric models to evaluate how productivity variations influenced wage adjustments, providing insights into labour market dynamics that could inform agricultural labour management practices.

D. Makwana et al. [21] analysed rainfall characteristics and moisture availability indices for crop planning in semi-arid regions of North Gujarat. They used machine learning algorithms to predict moisture availability and optimize crop planning schedules, ultimately enhancing agricultural productivity in water-scarce areas.

K. Ly et al. [22] employed geostatistical interpolation techniques to predict daily rainfall at the catchment scale in Belgium. By applying various variogram models, they achieved high-precision rainfall estimates, which are crucial for effective water resource management and agricultural planning.

Z. Jinglin et al. [23] used Support Vector Machine (SVM) and Particle Swarm Optimization (PSO) algorithms to predict precipitation data. Their research demonstrated that the PSO-SVM model provided superior accuracy in rainfall prediction compared to traditional methods, highlighting its potential application in agriculture.

T. Abebe and M. Endalie [24] developed artificial intelligence models for predicting monthly rainfall without relying on climatic data for meteorological stations in Ethiopia. Their innovative approach utilized machine learning algorithms to generate reliable rainfall forecasts, aiding in agricultural decision-making processes.

The contribution of accurate rainfall estimation is crucial for stakeholders in planning and increasing their production. With the world population projected to reach 9.2 billion by 2050, a 70% increase in global food production is necessary to meet the growing demand. Agriculture, being the cornerstone of human evolution, plays a vital role in feeding urban populations. To enhance agricultural management practices, modern techniques such as computer-aided plantation and crop monitoring have been introduced. These advanced methods, including precision agriculture and data-driven decision-making, are essential for improving crop yields and mitigating the impact of unpredictable factors like rainfall on crop production [25]. Rainfall, a critical variable affecting crop growth, can have both positive and negative impacts on agriculture. While adequate rainfall is essential for plant growth, excessive precipitation can lead to crop damage, especially in direct-seeded fields like rice crops. Farmers need to be vigilant and take precautionary measures to protect their crops from the adverse effects of overwhelming precipitation. Rainfall prediction models, incorporating machine learning techniques, have emerged as valuable tools for providing early warnings to farmers, enabling them to manage their agricultural lands efficiently and minimize crop damage [24]. Improved rain prediction models not only aid in safeguarding crops but also contribute to the optimal utilization of water resources. By utilizing sensory inputs from crop fields and historical rainfall data, machine learning approaches can accurately forecast rainfall patterns. These predictions are crucial for agronomists and farmers to plan their agricultural activities effectively, ensuring sustainable crop production and food security. Therefore, the development and implementation of reliable rainfall prediction models are essential for enhancing agricultural productivity and managing the impact of climate change on food security [26]. In conclusion, the integration of advanced agricultural techniques with precise rainfall prediction models can significantly benefit stakeholders in estimating and preparing for rainfall variations, ultimately leading to increased agricultural production and food security in the face of a growing global population.

The related work is summarized in Table 1.

Table 1

Summarized related work

Citation	Publication Year	ML Models	Performance measures used	Research Main Focus
[2]	2023	Multilayer Perceptron	Spearman's correlation coefficient	Rainfall prediction for rice crops
[3]	2022	Support Vector Regression, Lasso Regression, and Multiple Linear Regression	mean absolute error	Devastation of crops and rainfall prediction
[4]	2019	XGBoost	Accuracy, Kappa	precipitation prediction
[5]	2018	ARIMA, Support Vector Regression and ANN	Root Mean Squared Error	Rainfall Prediction

[6]	2018	Non-ML based approaches	Precipitation expectancy in mm	Region suitability for crops
[7]	2020	Crop growth model of FAO/IIASA	Difference between their estimation and actual rainfall	Study on Rainfall Patterns
[8]	2019	Statistical Models	Difference between total production model and yield model	Effect of flood on rice crop
[9]	2019	Logistic Regression, k-Nearest Neighbors, Decision Tree	Accuracy, Precision, Recall, F1, AUC	Rainfall prediction
[10]	2019	Deep Neural Network	Root mean squared error and normalized RMSE	Rainfall prediction
[11]	2021	Threshold based regression model	Standard error, rice production in tons	Flood impact on rice crops
[12]	2021	twelve-member convection-permitting ensemble (CPM) ensemble	Percentage relative change and areal mean intensity	Climate model for rainfall prediction in summers
[13]	2020	Extreme Value Theory (EVT), SARIMA	Normality Assumption Test, Significance Parameter Tests	Rainfall prediction
[14]	2020	VARMA statistical forecast method, ordinary Least Squares, robust linear regression model	Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF)	Temperature and rainfall predictions for rice crops
[15]	2021	Artificial Neural Networks and Partial method of least squares Regression (PLSR and BPNN)	RMSE	Rainfall prediction for rice yield in East China
[16]	2021	ANN and SVM	MSE R MAPE (%) Nash number RSR BIAS	Climate variables and rice yield
[17]	2021	Regression analysis	ANOVA	Rainfall prediction
[18]	2023	Statistical Models	Photosynthetic rate and transpiration rate	Plant responses to irrigation and rewatering
[19]	2021	Statistical Methods	Pettit's, the von Neumann ratio (VNR), Buishand's range (BR) and standard normal homogeneity (SNH) plus trend analysis Mann-Kendall (MK)	Rainfall prediction in Ethiopia
[20]	2006	Survey and statistical analysis	Relationship between rainshock and crop yield, probability density function	Productivity shocks within agricultural settings
[21]	2021	Statistical Model based on moisture levels	Probability density functions, chi-squared, rainfall probability	Rainfall prediction
[22]	2011	geostatistical interpolation techniques	Frequency percentage and RMSE	Rainfall prediction in Belgium
[23]	2017	Support Vector Machine (SVM) and Particle Swarm Optimization (PSO) algorithms	Accuracy and Run time consumed	Precipitation and rainfall prediction
[24]	2023	Adaptive Neuro-Fuzzy inference system ANFIS neural networks	Root Mean Square Error (RMSE), Nash–Sutcliffe model efficient coefficient (E), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and coefficient of determination (R2)	climatic data for meteorological stations in Ethiopia and rainfall prediction
[25]	2024	Statistical models, NetErosividade	R-factor, Spatiotemporal distribution of	rainfall erosivity estimation models and space–time distribution
[26]	2021	Pearson correlation technique, XGBoost, (Multivariate Linear Regression, Random Forest	Root mean squared error and Mean absolute Error methods	Rainfall Prediction

All the various research discussed above either implemented only one model or stuck to a limited number of prediction features. Our work tries to incorporate numerous environmental factors for rainfall prediction. Hence it explores suitability of ML algorithms for rainfalls prediction with more accuracy taking into account various weather features.

3. System Description

This section describes the steps carried out for the evaluation of the machine learning algorithms. The details about used dataset, pre-processing including missing treatment, feature selection is described. Reported results are based on 5-fold cross validation. Different Machine Learning algorithms which are frequently used in the relevant research domain were applied and comparison was drawn among their results. The performance evaluation is determined in the form of accuracy, precision, recall and mean squared error posing the problem as binary classification whether rainfall is expected in the next day or not.

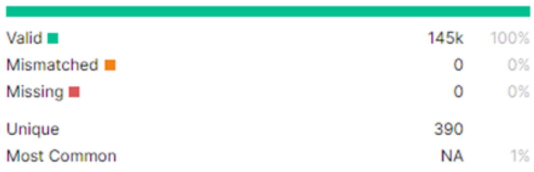
3.1 Dataset Description

The dataset used for this work contains the weather observations from various Australian weather stations. The dataset consists of 145460 instances and 24 attributes of which 11 attributes have been selected for rainfall prediction as independent variables/features. These selected 11 attributes were used for rainfall prediction selected as the independent variables. These selected dataset features are listed and described in Table 2. The dataset can be accessed and downloaded from [27]. Fig 1 sheds light on the dataset characteristics including percentage of missing values and null values. Fig. 2 summarizes the overall trends observed in the selected features pictorially.

Table 2

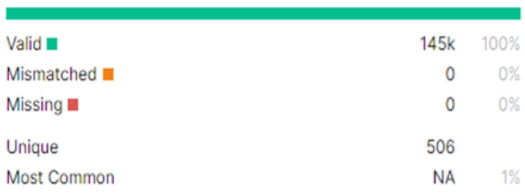
Dataset description according to attributes of interest, the dataset contains 12 attributes related to temperature, sunshine, evaporation rate, temperature and humidity recorded at 9 am, prior rainfall record in mm and the target feature RainTomorrow

Attributes	Description	Attributes	Description
MinTemp	minimum temperature °C	Temp9am	Temperature °C at 9 am
MaxTemp	maximum temperature °C	WindSpeed9am	Mean wind speed km/hr. per 10 min prior to 9 am
Rainfall	Recorded rainfall in day mm	Humidity9am	percentage of humidity at 9am
Evaporation	Class A pan evaporation mm in in a day measured at 9 am	Pressure9am	atmospheric pressure hpa computed to mean sea level at 9 am
Sunshine	No. of hrs. of bright sunshine in day	Cloud9am	clouded sky fraction at 9am.
WindGustSpeed	strongest wind gust speed km/h :24 hours to midnight	RainTomorrow	The dependent variable



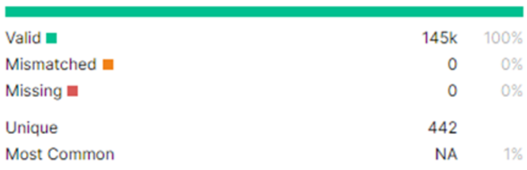
Valid 🟩	145k	100%
Mismatched 🟧	0	0%
Missing 🟥	0	0%
Unique	390	
Most Common	NA	1%

MinTemp



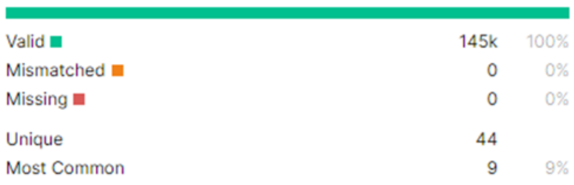
Valid 🟩	145k	100%
Mismatched 🟧	0	0%
Missing 🟥	0	0%
Unique	506	
Most Common	NA	1%

MaxTemp



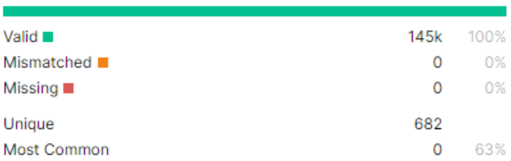
Valid 🟩	145k	100%
Mismatched 🟧	0	0%
Missing 🟥	0	0%
Unique	442	
Most Common	NA	1%

Temp9am



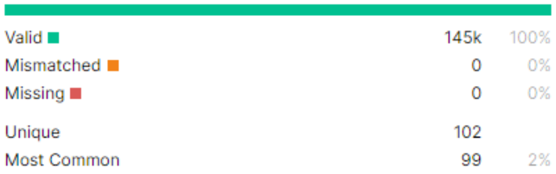
Valid 🟩	145k	100%
Mismatched 🟧	0	0%
Missing 🟥	0	0%
Unique	44	
Most Common	9	9%

WindSpeed9am



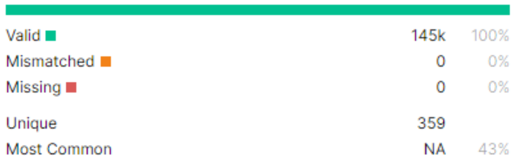
Valid 🟩	145k	100%
Mismatched 🟧	0	0%
Missing 🟥	0	0%
Unique	682	
Most Common	0	63%

Rainfall



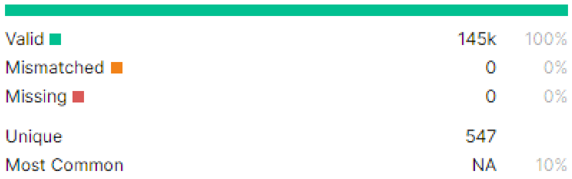
Valid 🟩	145k	100%
Mismatched 🟧	0	0%
Missing 🟥	0	0%
Unique	102	
Most Common	99	2%

Humidity9am



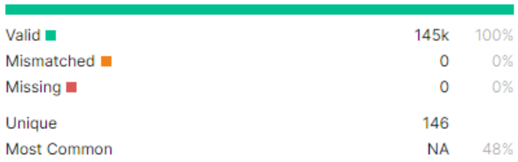
Valid 🟩	145k	100%
Mismatched 🟧	0	0%
Missing 🟥	0	0%
Unique	359	
Most Common	NA	43%

Evaporation



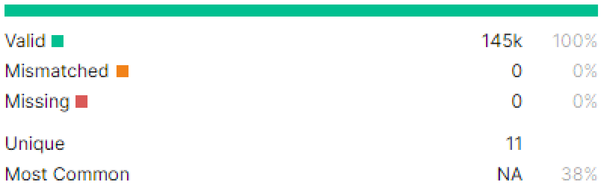
Valid 🟩	145k	100%
Mismatched 🟧	0	0%
Missing 🟥	0	0%
Unique	547	
Most Common	NA	10%

Pressure9am



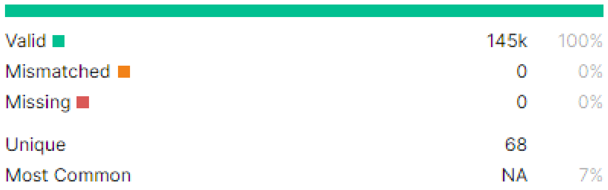
Valid 🟩	145k	100%
Mismatched 🟧	0	0%
Missing 🟥	0	0%
Unique	146	
Most Common	NA	48%

Sunshine



Valid 🟩	145k	100%
Mismatched 🟧	0	0%
Missing 🟥	0	0%
Unique	11	
Most Common	NA	38%

Cloud9am



Valid 🟩	145k	100%
Mismatched 🟧	0	0%
Missing 🟥	0	0%
Unique	68	
Most Common	NA	7%

WindGustSpeed

Fig. 1. Characteristics of Selected 11 Features [27], In Terms Of Details About Validation of Records, Missing Values, Number of Unique Values and The Most Frequent Value Per Attribute/Feature

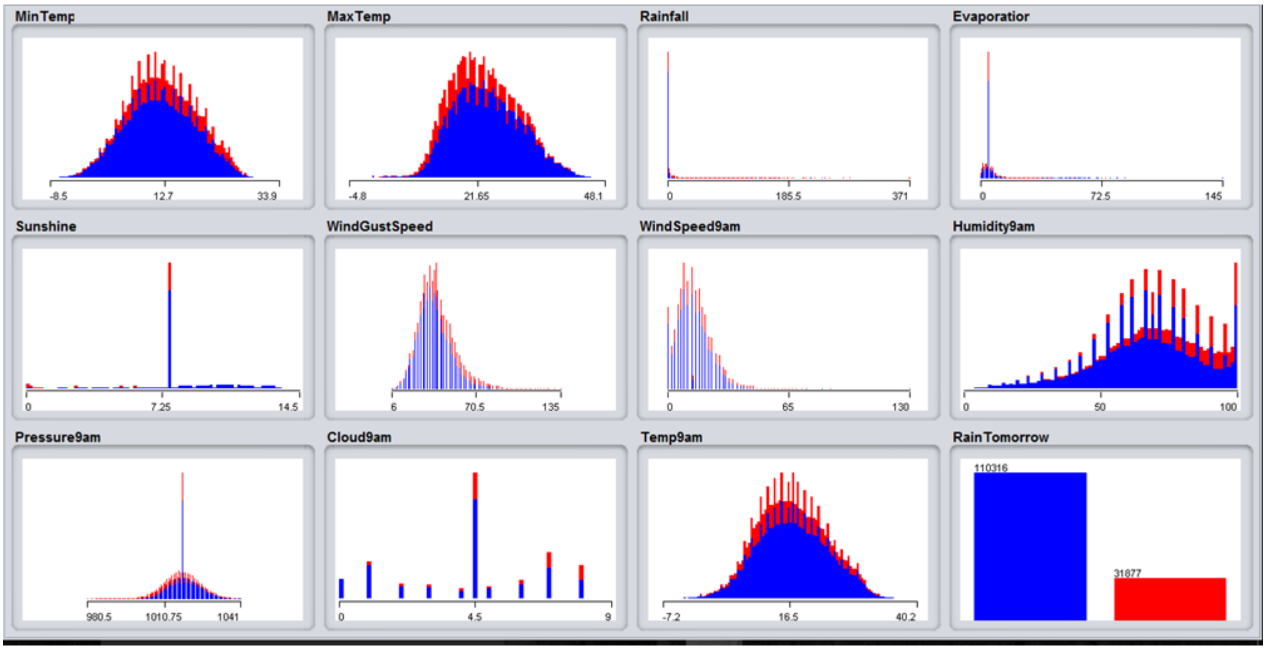


Fig. 2. Visualization of the Characteristics of Selected Features, Describing the Minimum and Maximum Values Per Attribute Along with The Spread and Frequency of Occurrence of a Certain Value

3.2 Data Pre-processing

Data pre-processing is utilized to convert the data into a more meaningful and useful format. It can include missing value treatment, conversion of values from one form to another, scale conversion, and feature selection. The following pre-processing steps were performed for data used for our work.

3.2.1 Missing values

There are several situations where missing values are inevitable in the observations of the dataset. Data cleaning and gleaning play a crucial role how well a model learns from the training data. There are many instances in the dataset that contained null values. For missing value treatment, we replaced the null values with their respective mean values attribute wise.

3.2.2 Feature selection

Feature Selection is employed for data dimension reduction. It can be performed automatically or manually. The purpose of feature selection is two-fold. Firstly, it reduces attribute count/dimension reduction. Secondly only those features are included that contribute the most to the dependent variable avoiding too much clutter helping the classifier decide in an effective and efficient manner. When there are many irrelevant characteristics/attributes in the data, this can reduce the precision of the prediction models and render the characteristics supported by the model irrelevant. Therefore, important /relevant feature selection can improve the accuracy as well as reduces training time. The dataset contained 24 features in which we have selected the most important 11 features such as minimum temperature, maximum temperature, rainfall, evaporation, sunshine, wind gust speed, wind speed at 9 am, humidity at 9 am, pressure at 9 am, clouds at 9 am, and temperature at 9 am. This feature selection was aided by employing Random Forest with 90 trees and 4 random features stopping the tree depth at 512 maximum splits. Random forest was used to compute predictor importance which resulted in selection of most important 11 features which can help improve prediction accuracy.

3.3 System Workflow

The input to the model is the dataset's 11 selected features and the output of the model is calculated in the form of accuracy and errors of prediction of the models. The task is treated as binary classification of whether there will be rainfall or not. For comparative evaluation WEKA API was used. The dataset was processed using 5-fold cross validation and with 5 repetitions of the algorithm. The presented results are the average of the performance measures of 5 repetitions of the algorithm evaluation. The number of instances used for training and testing were 113754, 28439 respectively. The following machine learning algorithms were evaluated: Naïve Bayes, LogitBoost, RIPPER, Decision Stump, AdaBoost, Random Forest, Artificial Neural network (ANN), and K*.

3.3.1 Evaluation measures

The evaluation was based on the following performance measures: accuracy, precision, recall, F1-measure, area under receiver operating characteristic curve (ROC curve), elapsed training time (average training time per sample) and elapsed testing time (average testing time per sample). On the basis of these performance measures as mentioned above, the best algorithm among them is chosen. Performance measures formulae are provided in Eq. (1) to (6) in terms of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). A complete workflow diagram (Fig. 3) is given below for a clearer understanding of the working.

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (1)$$

$$Precision = \frac{TP}{(TP+FP)} \quad (2)$$

$$Recall = \frac{TP}{(TP+FN)} \quad (3)$$

$$F1 - Measure = \frac{(2*TP)}{(2*TP+FP+FN)} \quad (4)$$

$$\text{Elapsed training time} = \frac{\text{(time consumed for training classifier)}}{\text{total number of samples in training subset}} \quad (5)$$

$$\text{Elapsed training time} = \frac{\text{(time consumed for testing classifier)}}{\text{total number of samples in testing subset}} \quad (6)$$

3.4 Applied Models

Numerous classifiers were evaluated, each belonging to a different category such as tree-based, bagging, boosting, lazy, rule based and Ensemble learning. The top five performing classifiers are briefly described in sections 3.4.1 to 3.4.5.

3.4.1 Random forest

Random forest or random decision forest is an ensemble learning technique for classification and regression. It operates by constructing a large number of decision trees using bagging. During training a user defined number of decision trees, are constructed with f random features for tree growth and pruning. The third hyper parameter for random forest is the depth of the tress controlled by maximum number of splits allowed. Due to multiple decision trees with f random features for tree split, random forest avoids the overfitting issue encountered in decision trees. The suitability of a certain classifier is dictated by the nature of data observations. However, random forest proves to have higher generalization ability and accuracy for many different sorts of data observations which was also the case in rainfall forecast.

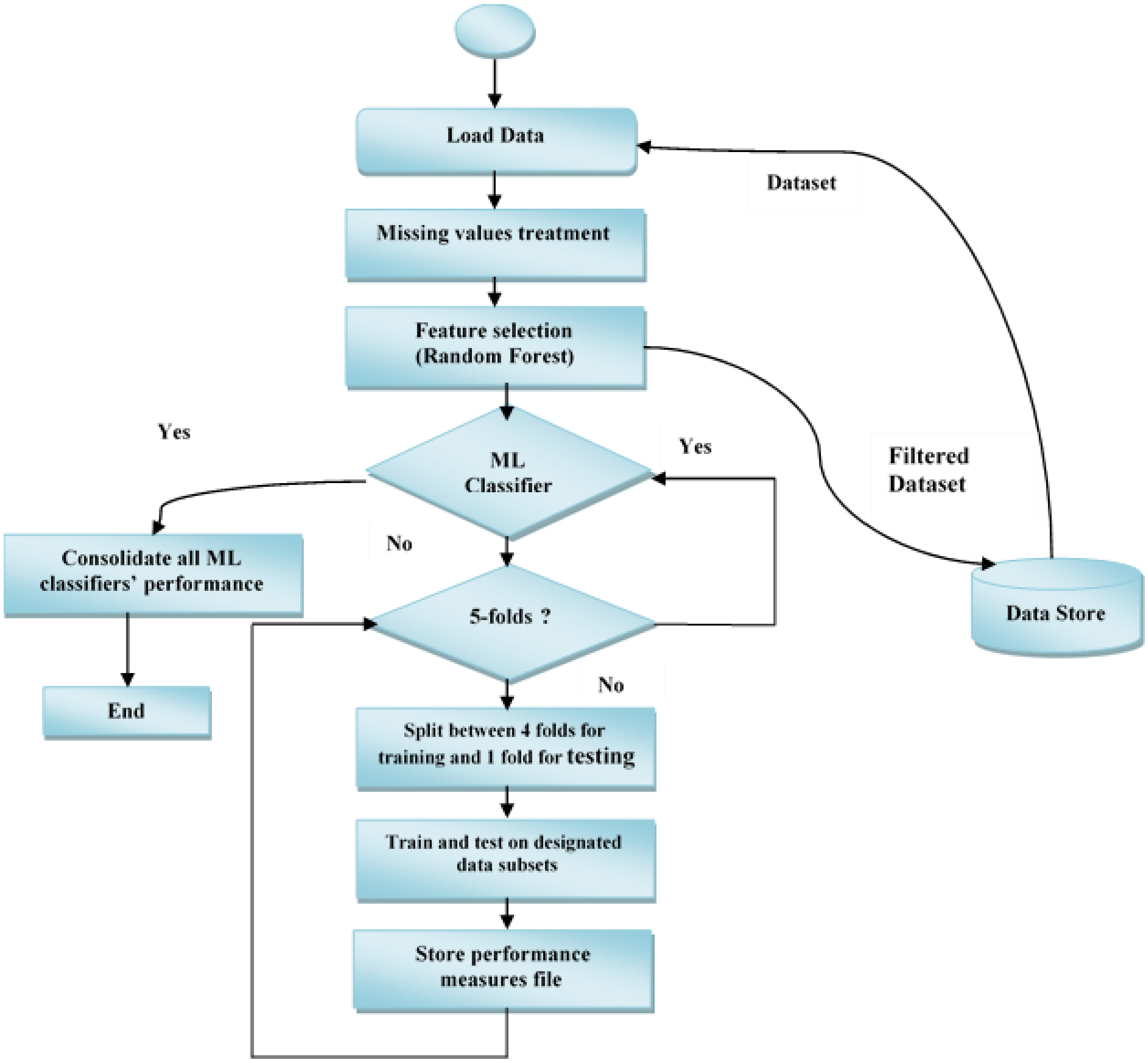

Load Data
Dataset
Missing values treatment
Feature selection (Random Forest)
Yes
ML Classifier
Yes
Filtered Dataset
Consolidate all ML classifiers' performance
No
5-folds ?
Data Store
End
No
Split between 4 folds for training and 1 fold for testing
Train and test on designated data subsets
Store performance measures file

Fig. 3. Workflow Diagram, Describing the Overall Functionality and Data Flow of the System by Incorporating Feature Selection and Then Best Performing Classifier Cascaded

3.4.2 Artificial neural networks

The artificial neural network (ANN) is essentially a brain simulation attempt. Neural network theory is based on the notion that certain key properties of biological neurons can be extracted and applied to simulations, creating a simulated and simplified brain. ANN employ the following three basic concepts: 1) connection force, 2) inhibition/excitation and 3) transfer function. Artificial neurons/nodes imitating human brain neurons are interconnected and the strength of this connection is normally assigned a numeric value between -1.0 for maximum inhibition and +1.0 for maximum excitation. Neurons are usually arranged as an input layer, an output layer, and a layer or layers hidden between the input and output layers. For rainfall prediction problem, back-propagation was utilized as learning method. The number of input neurons correspond to the number of 11 features of our dataset and two neurons in the output layer for prediction of either rainfall is predicted or not. Number of hidden layers and the number of neurons per hidden layers were varied and the reported results are based on the average of the computed results. The advantage of ANN is good accuracy, generalization ability, and very small prediction time. As once trained, the calculations for prediction are simple mathematical expressions. However, there are few drawbacks too such as many hyper parameters to be tuned, lots of time consumed during training, and possibility of overfitting.

3.4.3 LogitBoost

Simple regression functions are used as a basis for adaptation of learners to logistical models. The algorithm used behind is LogitBoost. The optimal number of iterations is cross-validated, which automatically selects the attributes. LogitBoost could be an analytics model that, in its fundamental form, uses a logistic function to model a binary variable, despite the fact that there are more complex extensions. LogitBoost (or logit) serves to model the probability of a certain class. or when a nominal variable with two values and a measuring variable is available such as success/failure, male/female, deceased/living, healthy/ill etc. LogitBoost regression is similar to rectilinear regression, except that the variable is nominal, rather than a measure. This will be utilized to show some categories of events such as deciding if an image contains a cat, puppy, lion, etc. Each object being recognized inside the picture would be assigned a probability between 0 and 1, with a sum of 1. Logistic regression is used to shed light on the data and clarify the relationship between a dependent binary variable and one or more independent factors on the nominal, ordinal, range or ratio level. Rainfall related climatic variables such as maximum and minimum temperature, evaporation and, morning and afternoon humidity etc were fed to the algorithm with varying hyper parameters and the averaged-out results are presented. The cross-validation analysis also shows that the logistic regression not only adequately fits the rainfall data which are utilized in the fitting procedure, it is often very successful in predicting rainfall for the longer-term data.

3.4.4 RIPPER

RIPPER (Repeated Incremental Pruning to Produce Error Reduction) as the name suggests is propositional rule learner. This algorithm in an enhanced and optimized version of IREP algorithm. This works basically in two stages. First, is the building stage and second one is the optimization stage. In the building stage, repeatedly rules are greedily grown and pruned until the stopping criteria is met. In the optimization stage, after generation of the initial rule set, two variations of each rule are pruned based on randomised data. A variant is generated based on an empty rule whereas the other is created by greedily including antecedents to the original rule.

3.4.5 Naive bayes

Naive Bayes allocates class labels to data instances, in which class labels are extracted from a limited set. This is a strong and unrealistic assumption when it comes to real data.; however, the technique is highly efficient for a wide range of complex problem domains. Naive Bayes is a method for predicting the probability of different classes depending on different attributes. No single algorithm is available to prepare these classifiers, but an algorithmic family supports a typical principle.: all Naïve Bayes classifiers assume that the value of a specific feature is independent of the value of the other feature, considering the category variable.

4. Simulation and Results

Naïve Bayes, LogitBoost, RIPPER, Decision Stump, AdaBoost, Random Forest, Artificial Neural network, and K* were the classifiers used for rainfall prediction in this research work. Our intention was to predict the rainfall, whether it will be occurring tomorrow or not, so that this prediction can help in the protection of rice crops from heavy rainfall. Numerous ML classifiers were evaluated using 5-fold cross validation according to the aforementioned performance measures and the ones who showed promising results are presented in this paper. Among all these algorithms, Random Forest was the best performing classifier that predicted with much better accuracy than other models.

Random Forest was the top performing classifier with maximum accuracy of 83.2%, followed by ANN (82.5%), LogitBoost (82.2%), RIPPER (82%), naïve Bayes (80.3%), AdaBoost (80.2%), and K*(79.2%) respectively in descending order. Evaluated performance measures are summarized in Table. 3.

Table 3

Summarization of performance measures by various ML classifiers in terms of accuracy, precision, recall, F1 score, average elapsed training time per sample in milliseconds, and average elapsed testing time per sample in milliseconds

Classifier	Accuracy	Precision	Recall	F1-measure	Elapsed training time (ms)	Elapsed testing time(ms)
Naïve Bayes	80.3	84.5	91.5	87.8	0.304	0.186
LogitBoost	82.2	84.4	94.5	89.2	3.51	0.0402
RIPPER	82	84.6	93.9	89	518	0.0396
Decision Stump	77.6	77.6	100	87.4	0.573	0.0144
AdaBoost	80.2	82.6	94.7	88.3	8.88	0.037
Random Forest	83.2	85.4	94.5	89.7	61.1	2.54
ANN	82.5	85	94	89.3	18.2	0.053
K*	79.2	84.7	89.2	86.9	0.015	8170

Table 3. highlights the best performing classifier according to each performance evaluation parameter in boldface. Whereas, the least performing one is depicted with red colour. Fig. 4 compares the accuracies in percentage obtained by different classifiers with Random Forest providing the maximum accuracy and decision stump at the bottom of this comparison. Fig. 5-8 summarizes comparison in precision, recall and F1-measure respectively. Again, maximum precision and F1-measure value was achieved by Random Forest and minimum by decision stump (Fig. 5 and 7 respectively). However, Fig. 6 shows that maximum recall was provided by Decision Stump and minimum by K*.

Fig. 8 and 9 shed lights on training and testing time consumed in milliseconds. Random Forest training consumed the maximum time but providing maximum accuracy. K* being a lazy approach consumed minimum training time but maximum testing time indicating that it is not suitable for real time prediction systems. Decision stump provided maximum 100 percent recall, but the accuracy was the minimum among reported classifiers.

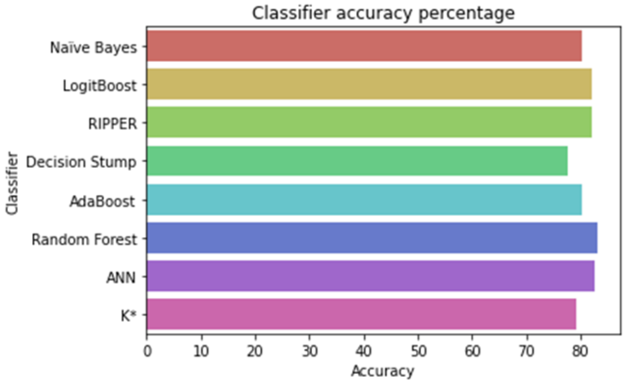


Fig. 4. Accuracy Percentage Obtained by Evaluated Classifiers, K*, ANN, Random Forest, AdaBoost, Decision Stump, RIPPER, LogitBoost and Naïve Bayes, Random Forest Clearly Stands as The Winner Followed by ANN and LogitBoost

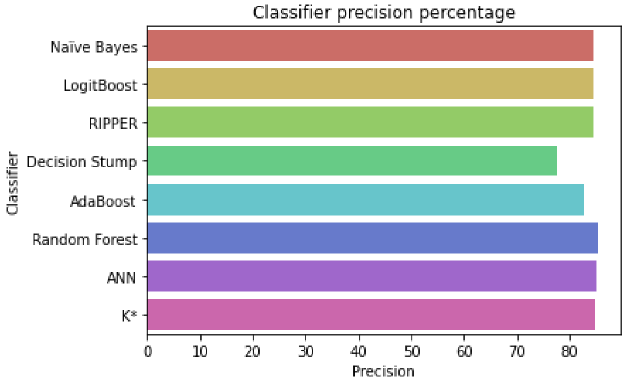


Fig. 5. Precision Percentage Obtained by Evaluated Classifiers, K*, ANN, Random Forest, AdaBoost, Decision Stump, RIPPER, LogitBoost and Naïve Bayes. Random Forest Showed Maximum Precision Followed by ANN and Then K*

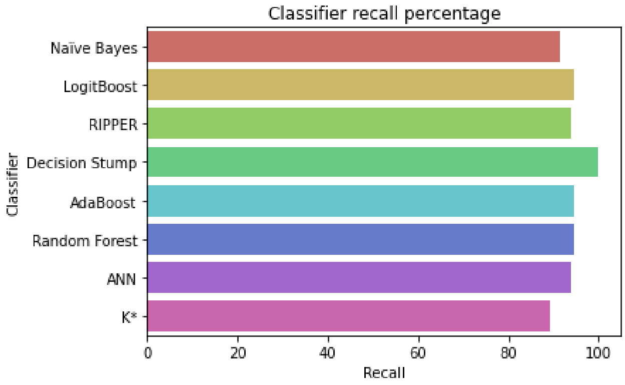


Fig. 6. Recall Percentage Obtained by Evaluated Classifiers, K*, ANN, Random Forest, AdaBoost, Decision Stump, RIPPER, LogitBoost and Naïve Bayes. In Terms of Recall, Decision Stump Showed Best Performance Followed by AdaBoost and Then Random Forest

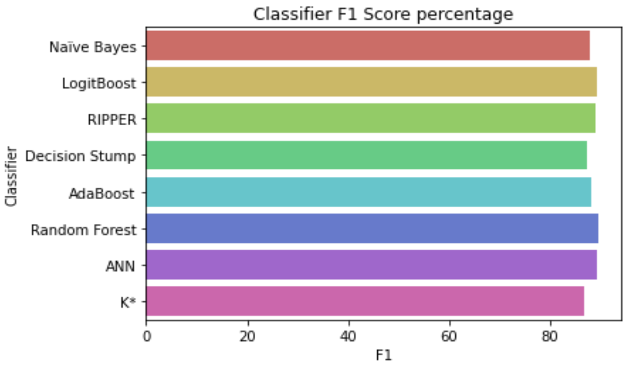


Fig. 7. F1 Percentage Obtained by Evaluated Classifiers, K*, ANN, Random Forest, AdaBoost, Decision Stump, RIPPER, LogitBoost and Naïve Bayes. F1 Score Performance Competition Was Again Dominated by Random Forest

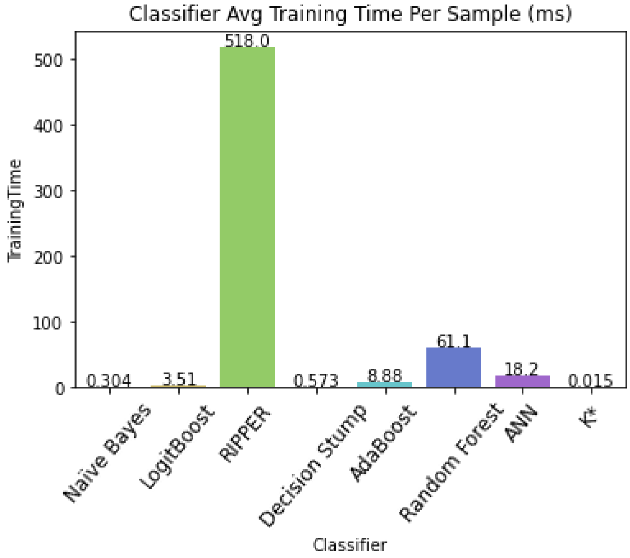


Fig. 8. Average Training Time Per Sample in Milliseconds Consumed by Classifiers, K*, ANN, Random Forest, AdaBoost, Decision Stump, RIPPER, LogitBoost and Naïve Bayes. RIPPER Showed Worst Performance In Terms Of Average Training Time Per Sample, K* and Naïve Bayes Had the Minimum Training Time Per Testing Sample In Millisecond

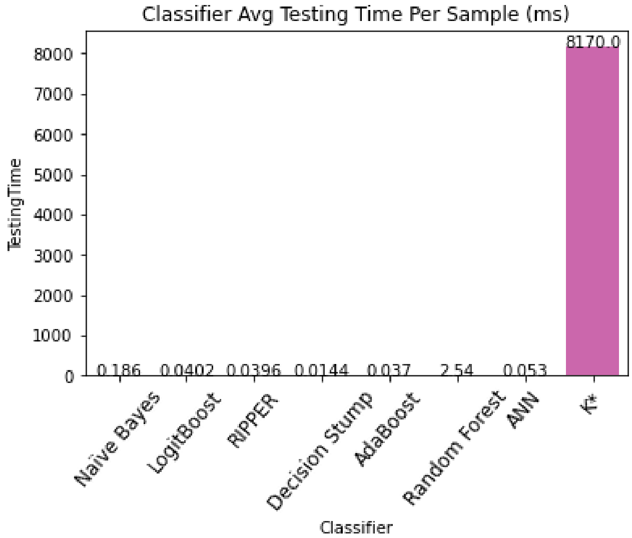


Fig. 9. Average Testing Time Per Sample in Milliseconds Consumed by Classifiers, K*, ANN, Random Forest, AdaBoost, Decision Stump, RIPPER, LogitBoost and Naïve Bayes. K* Had Minimum Average Training Time but Worst Testing Time

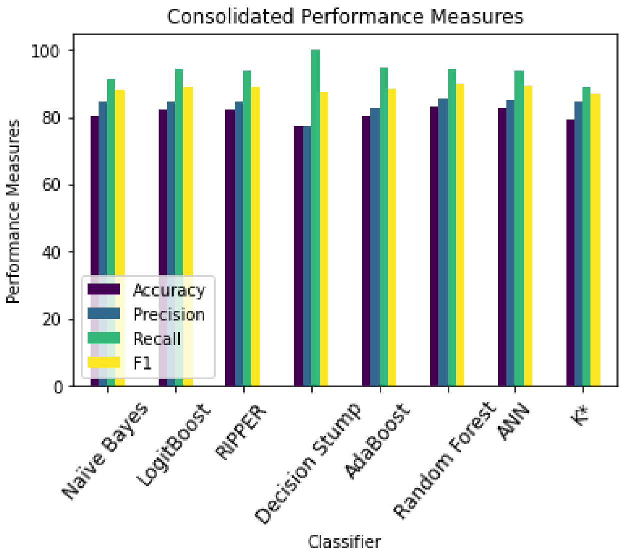


Fig. 10. Summarization of Classifier Performance, K*, ANN, Random Forest, AdaBoost, Decision Stump, RIPPER, LogitBoost and Naïve Bayes. When Accuracy, Precision, Recall and F1 Score Are Overall Compared, Random Forest Showed the Best Performance Followed By ANN and hen LogitBoost

All performance measures other than related to training and testing are summarized in Fig. 10.

It is evident from the results that Random Forest renders itself to be most suitable classifier for rainfall prediction, with 83.2 accuracy, 85.4 precision, 94.5 recall, and 89.7 F1 score. However, it consumed on average 61.1 ms during training per sample and 2.54 ms per sample during testing. If rainfall prediction system needs to be deployed in real time environment with high volume of predictions to make then 2.54 ms per prediction is much higher than 0.053 ms consumed by ANN and 0.0402 ms by LogitBoost. It must be noted that ANN and LogitBoost stand at 2nd and 3rd best performers with 82.5 and 82.2 accuracy respectively. Hence, in case high volume predictions where hard real time responses are needed then ANN or LogitBoost based system might be preferable than Random Forest based providing a good trade-off between swift response and classifier performance.

5. Conclusion and Future Work

Our prime objective was to protect the rice fields from heavy rainfall and our work was an attempt to predict precipitation under the combination of Machine Learning (ML) classifiers and predicting techniques. In this research, we have explored and successfully evaluated various ML classifiers for binary class prediction. The Random Forest classifier is thought to be a valuable and adaptive strategic solution for precipitation forecasting as its highest accuracy among all classifiers.

Also, we intend to install real world sensor equipped system in rice crop fields which by measuring the 11 features used in this study will be able to predict rainfall in real time and can generate alerts, alarms and messages for proper and prompt action. We also plan to incorporate time series analysis of the previous prediction to further improve the accuracy of rainfall prediction.

6. Limitations

There are few limitations of the proposed approach that should be considered for its real time application. First, the model's success significantly depends on the quality and amount of past meteorological data; any mistakes or missing data might lead to incorrect forecasts such as temperature and humidity recorded precisely at 9 am. Hence, required sensors must be deployed where ever the system needs to be deployed. Additionally, random forests, while showed best performance in rainfall prediction, its computational cost must be considered for real time application as the results revealed its average testing time per sample is much larger as compared to some other evaluated models such as ANN and LogitBoost.

7. References

[1] N. Alexandratos, “Expert meeting on how to feed the world in 2050”, Agric. Outlook, vol. 2050, no. 1, pp. 1–32, 2009.

[2] F. C. Eugenio et al., “Estimated flooded rice grain yield and nitrogen content in leaves based on RPAS images and machine learning”, F. Crop. Res., vol. 292, no. December 2022, p. 108823, 2023.

[3] D. Mahajan and S. Sharma, “Prediction of rainfall using machine learning”, 4th Int. Conf. Emerg. Res. Electron. Comput. Sci. Technol. ICERECT 2022, vol. 9, no. 01, 2022.

[4] R. Aguasca-Colomo, D. Castellanos-Nieves, and M. Méndez, “Comparative analysis of rainfall prediction models using machine learning in islands with complex orography: Tenerife Island”, Appl. Sci., vol. 9, no. 22, 2019.

[5] U. Shah, S. Garg, N. Sisodiya, N. Dube, and S. Sharma, “Rainfall prediction: Accuracy enhancement using machine learning and forecasting techniques”, PDGC 2018 - 2018 5th Int. Conf. Parallel, Distrib. Grid Comput., pp. 776–782, 2018.

[6] M. Rokonuzzaman, M. Rahman, M. Yeasmin, and M. Islam, “Relationship between precipitation and rice production in Rangpur district”, Progress. Agric., vol. 29, no. 1, 2018.

[7] S. Sujariya, N. Jongrungklang, B. Jongdee, T. Inthavong, C. Budhaboon, and S. Fukai, “Rainfall variability and its effects on growing period and grain yield for rainfed lowland rice under transplanting system in Northeast Thailand”, Plant Prod. Sci., vol. 23, no. 1, pp. 48–59, 2020.

[8] Hassan, S.M., Kling, R., Zahran, S. and Bayhem, J., 2019. Impacts of flooding on the rice production of Bangladesh: A panel data study (Doctoral dissertation, Masters dissertation, Colorado State University, Fort Collins, Colorado).

[9] N. Oswal, “Predicting rainfall using machine learning techniques,” 2019.

[10] M. H. Yen, D. W. Liu, Y. C. Hsin, C. E. Lin, and C. C. Chen, “Application of the deep learning for the prediction of rainfall in Southern Taiwan”, Sci. Rep., vol. 9, no. 1, pp. 1–9, 2019.

[11] M. Rahman, M. Islam, M. Rahaman, M. Sarkar, R. Ahmed, and M. Kabir, “Identifying the threshold level of flooding for rice production in bangladesh: An empirical analysis”, J. Bangladesh Agric. Univ., no. 0, p. 1, 2021.

[12] Y. Chen, A. Paschalis, E. Kendon, D. Kim, and C. Onof, “Changing spatial structure of summer heavy rainfall, using convection-permitting ensemble”, Geophys. Res. Lett., vol. 48, no. 3, pp. 1–12, 2021.

[13] E. R. M. Putri, I. G. A. Riska Astari, and N. Wahyuningsih, “Rainfall forecasting with climate change detection and its pattern relationship to rice production,” J. Phys. Conf. Ser., vol. 1490, no. 1, 2020.

[14] M. Bagbohouna, D. S. Ragatoa, S. O. Simon, and I. K. Edjame, “Rainfall and temperature predictions: implications for rice production in the lower river region of the gambia,” Univers. J. Agric. Res., vol. 8, no. 4, pp. 97–123, 2020.

[15] Y. Guo, H. Xiang, Z. Li, F. Ma, and C. Du, “Prediction of rice yield in east China based on climate and agronomic traits data using artificial neural networks and partial least squares regression”, Agronomy, vol. 11, no. 2, 2021.

[16] L. Wickramasinghe, R. Weliwatta, P. Ekanayake, and J. Jayasinghe, “Modeling the relationship between rice yield and climate variables using statistical and machine learning techniques”, J. Math., vol. 2021, 2021.

[17] A. District, B. Zone, and Y. Jambo, “Statistical Analysis of Rainfall Variability Effect on Wheat Production and Coping Strategies of Farmers: The Case of Agarfa District, Bale Zone, Ethiopia”, J. Biol. Agric. Healthc., vol. 11, no. 4, pp. 14–24, 2021.

[18] F. Müller et al., “Morphological and physiological responses of Calobota sericea plants subjected to water limitation and subsequent rewatering”, African J. Range Forage Sci., vol. 40, no. 2, pp. 141–158, 2023.

[19] Y. S. Getahun, M. H. Li, and I. F. Pun, “Trend and change-point detection analyses of rainfall and temperature over the Awash River basin of Ethiopia”, Heliyon, vol. 7, no. 9, p. e08024, 2021.

[20] S. Jayachandran, “Selling labor low: Wage responses to productivity shocks in developing countries”, J. Polit. Econ., vol. 114, no. 3, pp. 538–575, 2006.

[21] N. Resources, “planning in semi arid region of north Gujarat,” vol. 23, no. December, pp. 409–415, 2021.

[22] S. Ly, C. Charles, and A. Degré, “Geostatistical interpolation of daily rainfall at catchment scale: The use of several variogram models in the Ourthe and Ambleve catchments, Belgium”, Hydrol. Earth Syst. Sci., vol. 15, no. 7, pp. 2259–2274, 2011.

[23] J. Du, Y. Liu, Y. Yu, and W. Yan, “A prediction of precipitation data based on support vector machine and particle swarm optimization (PSO-SVM) algorithms”, Algorithms, vol. 10, no. 2, 2017.

[24] W. T. Abebe and D. Endalie, “Artificial intelligence models for prediction of monthly rainfall without climatic data for meteorological stations in Ethiopia”, J. Big Data, vol. 10, no. 1, 2023.

[25] L. Wang, Y. Li, Y. Gan, L. Zhao, W. Qin, and L. Ding, “Rainfall erosivity index for monitoring global soil erosion”, Catena, vol. 234, no. October 2023, p. 107593, 2024.

[26] C. M. Liyew and H. A. Melese, “Machine learning techniques to predict daily rainfall amount”, J. Big Data, vol. 8, no. 1, 2021.

[27] J. Young, “Rain in Australia”, Kaggle.Com, 2019. [Online]. Available: https://www.kaggle.com/datasets/jsphyg/weather-dataset-rattle-package.