Performance Comparison of Machine Learning Models in Rainfall Prediction: A Case Study of Mkwinda

by

Hobby Bwanali

PROJECT REPORT

Submitted to the Faculty of Agriculture in Partial Fulfilment of the Requirements for the Degree of

Bachelor of Science in Irrigation Engineering

Lilongwe University of Agriculture and Natural Resources (LUANAR)
Bunda College, Lilongwe, Malawi

August 2025

APPROVAL

The report committee for Hobby Bwanali, the Head of the Agricultural Engineering Department and the Dean of the Faculty of Agriculture Certify that this is the approved version of the following report:

Performance Comparison of Machine Learning Models in Rainfall Prediction: A Case Study of Mkwinda

APPROVED BY:

Supervisor:____

Dr. Chikondi Makwiza

Head of Department: ____

Prof. Christopher Kanali

Dean of Faculty: ____

Dr. Vincent Mgoli Mwale

DEDICATION

This work is dedicated to my loving mother E. Mwase and sister Agness, whose unwavering support and encouragement have been the foundation of my success. Their belief in my dreams have inspired me every step of the way.

ACKNOWLEDGEMENTS

I’m very thankful to my supervisor, Dr. C. Makwiza, for the helpful advice, support, and knowledge that really guided me through this research.

I also want to thank my friends Martin Chirwa, Joseph Kukacha, Joseph Chawinga, Joshua Munkhondya, and Holyce Phiri for encouraging me and keeping me motivated.

Lastly, I truly appreciate everyone who helped or supported this project in any way.

Above all, I thank God for blessing me with the strength, wisdom, and perseverance to complete this work.

ABSTRACT

Performance Comparison of Machine Learning Models in Localized Rainfall Prediction

by

Hobby Bwanali, BSc in Irrig. Eng.

Lilongwe University of Agriculture and Natural Resources (LUANAR)
Bunda College, 2025
SUPERVISOR: Dr Chikondi Makwiza

This study compares the performance of three machine learning models namely Logistic Regression (LR), Random Forest (RF), and Multi-Layer Perceptron (MLP) for predicting rainfall occurrence in Mkwinda, a rural area in Lilongwe District, Malawi. Historical and real-time weather data, including rainfall, temperature, and humidity, were collected and preprocessed to train the models. Each model was evaluated using classification metrics such as accuracy, precision, recall, F1-score, and ROC AUC. The results showed that while Logistic Regression achieved high recall, it produced many false positives. Random Forest offered improved precision but lower sensitivity. The MLP model outperformed the others with the highest accuracy (83.6%) and ROC AUC (0.8735), demonstrating its ability to handle complex, non-linear patterns in the data. These findings highlight the potential of neural networks in enhancing localized rainfall prediction and support the integration of advanced ML techniques into climate-sensitive decision-making.

Keywords: Rainfall Prediction, Machine Learning, Neural Networks.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

ANN	Artificial Neural Network
AUC	Area Under the Curve
CNN	Convolutional Neural Network
DCCMS	Department of Climate Change and Meteorological Services
FAO	Food and Agriculture Organization
FFNN	Feed Forward Neural Network
FN	False Negative
FP	False Positive
LR	Logistic Regression
MAE	Mean Absolute Error
ML	Machine Learning
MLP	Multi-Layer Perceptron
RF	Random Forest
RMSE	Root Mean Squared Error
ROC	Receiver Operating Characteristic
SVR	Support Vector Regression
TN	True Negative
TP	True Positive

1 INTRODUCTION

1.1 Background Information

Malawi's agricultural systems remain vulnerable to rainfall variability, a challenge magnified by climate change. Smallholder farmers, who form the majority of food producers, face increasing uncertainty in their planting and harvesting decisions due to unreliable seasonal forecasts. Accurate and reliable rainfall prediction is, therefore, essential for improving climate resilience and guiding agricultural practices in Malawi (FAO, 2020).

Traditionally, rainfall forecasting in Malawi has relied on weather predictions provided by the Department of Climate Change and Meteorological Services (DCCMS). While these systems are valuable, they sometimes struggle to provide accurate forecasts for specific local and remote areas due to the limited number of weather stations and gaps in weather data coverage. Additionally, traditional models can require expensive infrastructure and may not easily adjust to the unique weather patterns of smaller regions. As a result, there has been growing interest in data-driven approaches, such as machine learning (ML), which offer improved prediction capabilities, especially in areas where weather monitoring infrastructure is lacking.

Machine learning offers a promising alternative by extracting predictive signals from existing historical data. A 2023 study by Manjolo demonstrated this potential, applying logistic regression to predict rainfall in some areas in Lilongwe, Malawi with 50 - 80% accuracy using historical and real time meteorological data. While the study demonstrated that rainfall prediction was feasible using a basic classification model, its performance was limited by the simplicity of the algorithm and lack of exploration of alternative machine learning models. The results highlighted the potential of localized prediction but also revealed the need for further investigation using more advanced and diverse ML techniques. This current research builds upon that foundation by comparing the performance of multiple models, including both simple and more advanced approaches, to identify the most effective method for localized rainfall prediction.

This research specifically focuses on evaluating and comparing the predictive performance of three machine learning models Logistic Regression (simple linear), Random Forest (ensemble), and Multilayer Perceptron (neural network) using historical weather data collected from a local station.

1.2 Problem Statement

While machine learning (ML) models have been applied in weather forecasting at national and regional scale, their utility and accuracy for predicting rainfall in specific local or remote areas in Malawi remains largely unexamined. Many existing studies prioritize broad-scale applications, thereby overlooking the unique microclimatic variations that significantly influence localized rainfall patterns. This prevalent lack of comparative evaluations among different ML models for precise, localized rainfall prediction, particularly using historical meteorological data, constitutes a significant knowledge gap.

1.3 Justification

This study evaluates a data-driven approach to rainfall prediction using machine learning (ML) models trained and tested on historical and real time weather data specific to local areas. These ML techniques provide a cost-effective solution, enabling communities to access more precise forecasts, even in regions lacking advanced weather monitoring infrastructure.

By testing and comparing different ML models performance in a localized context, this research will contribute to the broader efforts of integrating machine learning into climate forecasting in developing countries like Malawi. The findings will not only inform the selection of suitable models for rainfall prediction at local and small scale but also support future research in similar low-resource settings. In the long term, improved localized predictions can help enhance agricultural planning, disaster preparedness, and climate adaptation strategies in vulnerable communities or areas where existing forecasting methods are unreliable.

2 OBJECTIVES

2.1 Main Objective:

To compare the performance of simple machine learning models, artificial neural networks (ANNs), and ensemble learning models for predicting rainfall occurrence.

2.2 Specific Objectives:

i. To Implement and train three machine learning models Logistic Regression, Random Forest, and Multi-Layer Perceptron for rainfall occurrence prediction.

ii. To compare the performance of the machine learning models using standard metrics.

iii. To determine the most suitable model for local use.

3 LITERATURE REVIEW

Rainfall prediction has been an important area of study because of its impact on farming, managing water supplies, and preparing for natural disasters. Recent progress in Machine Learning (ML) has improved weather forecasting by making it possible to find patterns in large sets of past weather data. This review looks at how ML models such as Logistic Regression (LR), Multi-Layer Perceptron (MLP), and Random Forest (RF) have been used in rainfall prediction, especially in specific local settings.

3.1 Machine Learning in Rainfall Prediction

Machine Learning techniques have been widely adopted for various prediction tasks due to their ability to handle large datasets and uncover hidden patterns. Studies have demonstrated the efficacy of ML models in improving prediction accuracy compared to traditional statistical methods. For instance, Abdullah & Said (2023) evaluated regression models such as Random Forest Regression (RFR) and Support Vector Regression (SVR) for rainfall prediction in Aligarh, India, highlighting the superior performance of ensemble methods. Similarly, Barrera-Animas et al. (2021) conducted a comparative analysis of modern ML algorithms, including LSTM networks and Gradient Boosting Regressors, for time-series forecasting, emphasizing the importance of feature selection and model optimization.

Recent studies have explored the use of ensemble learning techniques, such as K-Stars, for rainfall prediction. Tüysüzoğlu et al. (2023) proposed an ensemble model that combines multiple classifiers to improve prediction accuracy, achieving significant improvements in classification metrics. Additionally, Tüysüzoğlu et al. (2023) demonstrated the effectiveness of convolutional neural networks (CNNs) in enhancing precipitation estimation, particularly in regions with complex climatic conditions.

3.2 Evaluation Metrics for Rainfall Prediction Models

When assessing the performance of machine learning models, especially in critical applications like rainfall prediction, a variety of performance metrics are employed. These metrics quantify how well a model's predictions align with actual observations, providing insights into its accuracy, robustness, and suitability for specific tasks. For regression tasks, where continuous values are predicted (e.g., rainfall amount), common metrics include Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R^2). For classification tasks (e.g., predicting rainfall occurrence), metrics such as Accuracy, Precision, Recall, F1 Score, and tools like the Confusion Matrix and Receiver Operating Characteristic (ROC) Curve are frequently used.

Studies such as Abdullah & Said (2023) and Pan et al. (2019) have used these metrics to compare regression models, with ensemble methods frequently demonstrating higher R^2 and lower RMSE values. Similarly, Tüysüzoğlu et al. (2023) emphasized the importance of F1 Score for imbalanced datasets, where missed rainfall events (False Negatives) carry significant consequences.

Kumari et al. (2023) compared various machine learning models for rainfall prediction in urban metropolitan cities, evaluating classification metrics such as precision, recall, and F1 score.

3.2.1 Classification Performance Metrics

For models that predict whether rainfall will occur (a binary outcome), the following metrics are commonly used:

- **Accuracy:** Accuracy is the ratio of correctly predicted instances to the total instances in the dataset. It provides a general measure of how well the model performs across all classes.

- **Precision:** Precision is the ratio of correctly predicted positive observations (e.g., predicted rainfall that actually occurred) to the total predicted positives. It is useful for understanding the accuracy of positive predictions, especially in imbalanced datasets.

- **Recall:** Recall (also known as Sensitivity or True Positive Rate) is the ratio of correctly predicted positive observations to all observations in the actual class. It measures the model's ability to identify all relevant instances (e.g., all actual rainfall events).

- **F1 Score:** The F1 Score is the weighted average of Precision and Recall. It provides a balance between precision and recall, especially useful when the dataset has imbalanced classes.

- **Confusion Matrix and ROC Analysis**: The Confusion Matrix breaks down performance into True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN). This matrix helps analyze specific misclassification patterns in rainfall prediction.

  The ROC Curve plots the True Positive Rate (recall) against the False Positive Rate, helping select optimal probability thresholds. The AUC score quantifies the model’s ability to distinguish between rainfall and non-rainfall events.

3.3 Overview of ML Techniques in Rainfall Prediction

3.3.1 Logistic Regression (LR)

Logistic Regression is a simple yet interpretable model often used for binary classification tasks, including rainfall occurrence prediction. However, its performance is limited when dealing with complex, non-linear relationships in meteorological data. Studies have shown that while LR provides a baseline for comparison, it often struggles with the intricacies of atmospheric variables

Ajitha et al. (2023) compared Logistic Regression with advanced ML models, such as LightGBM and XGBoost, for rainfall prediction, highlighting the limitations of LR in handling large datasets. Similarly, Prottasha et al. (2023) evaluated LR alongside neural networks for short-term rainfall prediction, concluding that LR is better suited for simpler datasets.

3.3.2 Neural Networks for Rainfall Forecasting

MLP, a type of artificial neural network, has gained attention for its ability to model non-linear relationships and capture intricate patterns in data. Research by Song & Chen (2021) demonstrated the effectiveness of MLP in annual precipitation prediction, particularly when combined with decomposition methods. However, the model's reliance on extensive hyperparameter tuning and computational resources poses challenges for practical implementation.

Pan et al. (2019) explored the use of combined convolutional and long short-term memory (LSTM) networks for monsoon precipitation prediction, achieving superior results compared to standalone MLP models. Additionally, Hess & Boers (2022) applied deep learning techniques to improve numerical weather prediction, showcasing the potential of MLP in enhancing rainfall forecasts.

As shown in Figure 1, a MLP consists of an input layer, one or more hidden layers, and an output layer.

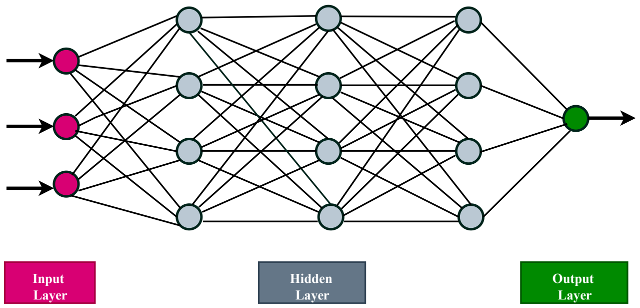

Figure 1. A MLP neural network with three hidden layers

The output (**y**) of a neural network can be mathematically represented as**:**

$$\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}) \tag{1}$$

Where: **W** = matrix of weights, **x** = vector of input features, **b** = bias term (a constant added to the output), **f** = activation function

3.3.3 Random Forest (RF)

Random Forest, an ensemble learning method, is renowned for its robustness and generalization capabilities. Studies have highlighted its balanced approach to handling both linear and non-linear relationships in rainfall prediction.

Hill et al. (2020) utilized Random Forest for forecasting excessive rainfall, demonstrating its effectiveness in capturing complex weather patterns. Furthermore, Abdullah & Said (2023) applied RF alongside CatBoost Regression for daily and monthly rainfall prediction, achieving strong correlations in their results.

3.4 Localized Rainfall Prediction

Localized rainfall prediction models are essential for addressing microclimatic variations that significantly impact agricultural planning and water resource management. Despite the advancements in ML, there is a scarcity of studies focusing on localized contexts, particularly in regions like Malawi. This gap underscores the need for tailored models that incorporate region-specific atmospheric variables and optimize prediction accuracy.

Tüysüzoğlu et al. (2023) emphasized the importance of localized datasets in improving prediction accuracy, using high-resolution weather data from local meteorological stations. Similarly, Pan et al. (2019) highlighted the role of feature selection in enhancing localized rainfall prediction models.

3.5 Emerging Trends in Rainfall Prediction

Recent advancements in ML have introduced hybrid models that combine the strengths of multiple algorithms. For instance, hybrid models integrating Long Short-Term Memory (LSTM) networks with Random Forest have shown promise in capturing both temporal dependencies and feature importance. Additionally, the use of satellite imagery and remote sensing data has enhanced the spatial resolution of rainfall predictions, enabling more precise forecasts.

Hess & Boers (2022) proposed physically constrained generative adversarial networks (GANs) for improving precipitation fields, demonstrating the potential of deep learning in enhancing rainfall prediction. Furthermore, Pan et al. (2019) explored probabilistic deep learning techniques for seasonal forecasts, achieving significant improvements in prediction accuracy.

3.6 Challenges and Future Directions with machine learning

While ML models have demonstrated significant potential, challenges remain in terms of data quality, computational requirements, and model interpretability. Future research should focus on integrating additional environmental predictors to enhance model accuracy. Moreover, the application of advanced optimization techniques, such as Bayesian optimization, can streamline hyperparameter tuning and improve model performance.

4 METHODOLOGY

4.1 Study Area

4.1.1 Description of the Study Area

The study focused on Mkwinda, a rural village located a short distance from Bunda College Campus in Lilongwe District in the central region of Malawi and home to the country's administrative capital, like most other Malawian communities, Mkwinda is primarily an agricultural community, where farming activities are highly dependent on rainfall. The community's reliance on agriculture highlights the importance of a reliable rainfall predictions to support farmers in making informed decisions. Figure 2 shows the locations of Mkwinda village, Lilongwe District.

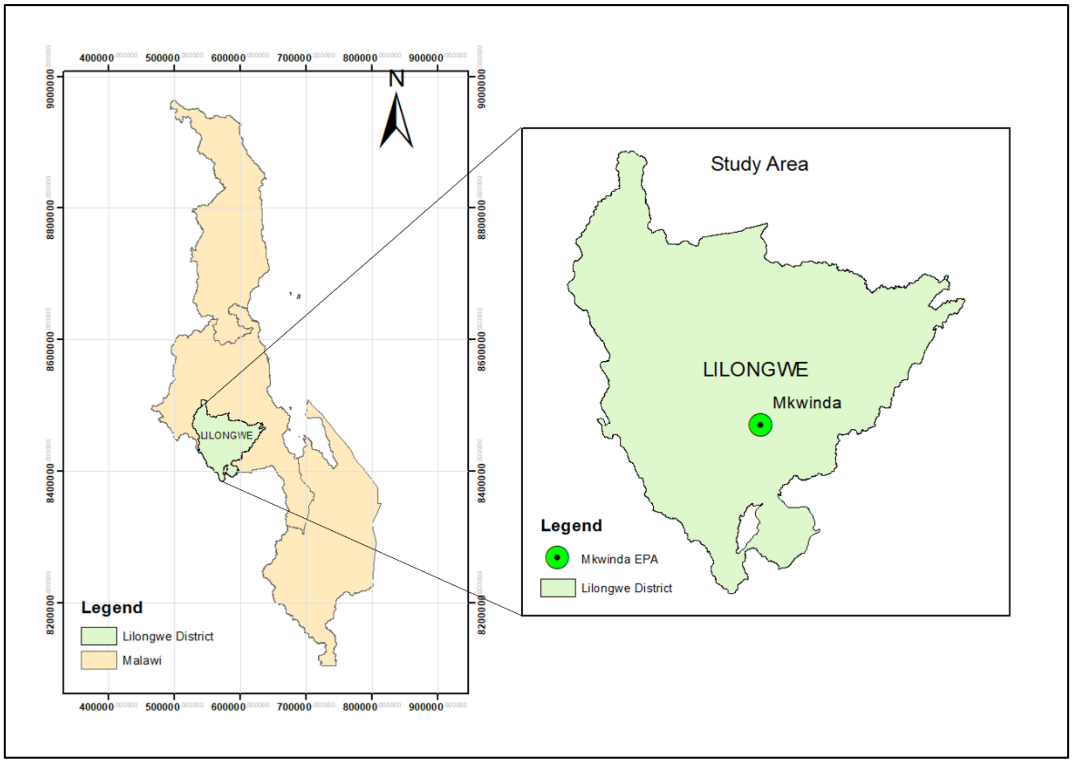

Figure 2. Map of study area

4.2 Machine Learning Flow Diagram for Rainfall Prediction

The machine learning flow diagram for rainfall prediction includes the following steps: Data Collection, Data Preprocessing, Feature Engineering, Model Training, Model Evaluation, Prediction. Figure 3 illustrates the overall architecture.

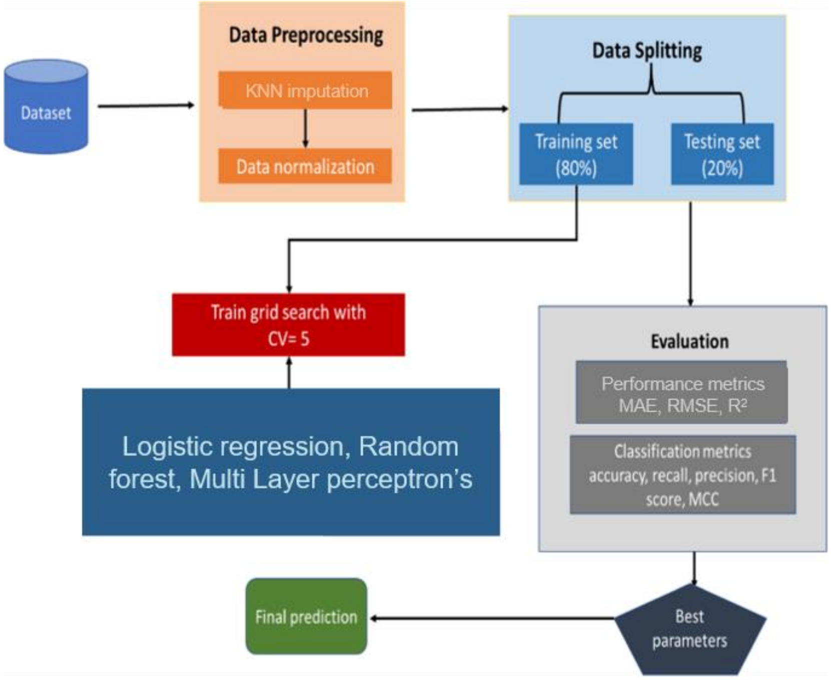
Dataset
Data Preprocessing
KNN imputation
Data normalization
Data Splitting
Training set (80%)
Testing set (20%)
Train grid search with CV= 5
Logistic regression, Random forest, Multi Layer perceptron's
Evaluation
Performance metrics MAE, RMSE, R²
Classification metrics accuracy, recall, precision, F1 score, MCC
Best parameters
Final prediction

Figure 3. Machine Learning Flow Diagram for Rainfall Prediction

4.3 Data Collection

Data collection involved obtaining both historical and real-time rainfall data. Historical rainfall data was gathered from various sources, including the crop science department at (LUANAR Bunda Campus), local agriculture offices (Mkwinda EPA) and online databases (NASA Power website). These datasets provided long-term records, which were useful for analyzing trends and training the machine learning models for prediction.

In addition to historical data, real-time rainfall data was collected using automated rain gauge systems developed with Arduino technology. These rain gauges were programmed to record cumulative rainfall at the specified intervals and were installed in the study area. At the agricultural office. Real-time data was essential for validating the models and ensuring their accuracy in predicting current and future rainfall patterns.

To complement rainfall data, additional weather parameters including temperature, and humidity were collected. These parameters were included because they have a significant influence on rainfall and could improve the predictive performance of the models. By incorporating both historical and real-time data, the study created a comprehensive dataset that reflected the unique weather conditions of the study area.

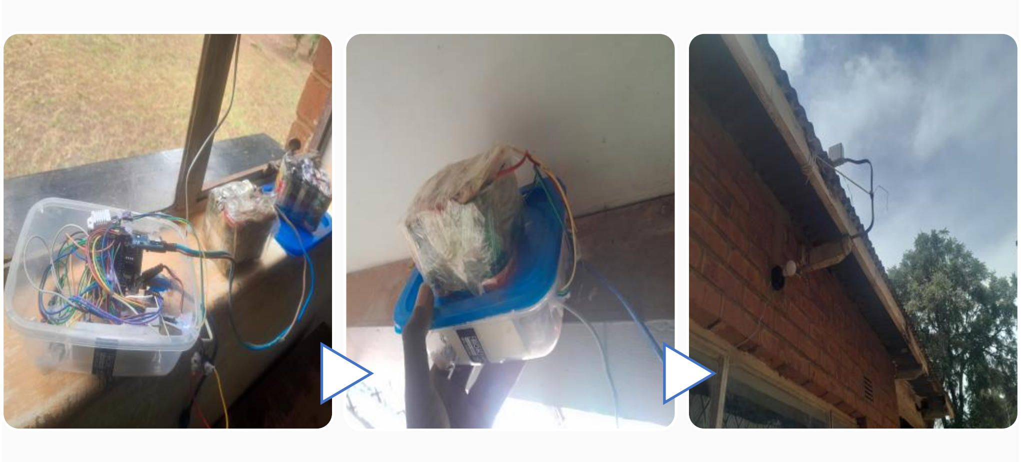

Figure 4. Real-time Rainfall Data Collection at Mkwinda

4.4 Data Analysis

Data exploration and analysis was necessary to ensure the reliability of the rainfall dataset used in this research. This phase aimed to clean, transform, and prepare the data for use in machine learning models while extracting meaningful patterns and insights. Python was the primary tool used for data analysis with its libraries like Pandas, NumPy, Matplotlib, and Scikit-learn. Microsoft Excel was also used for preliminary inspection, manual corrections, and quick visualizations to supplement the Python-based analysis.

4.4.1 Dataset Description

The dataset included the parameters presented in Table 1:

Table 1: Features descriptions of the train dataset

Variable	Unit	Description
rainfall	mm/day	Daily precipitation accumulation.
temp_max	°C	Maximum daily temperature.
temp_min	°C	Minimum daily temperature.
s_humidity	%	Specific humidity (mass of water vapor per unit air mass).
r_humidity	%	Relative humidity (water vapor present vs. maximum possible at a given temp).
wind_speed_max	m/s	Peak wind speed during the day.
wind_speed_min	m/s	Minimum wind speed recorded.
will_rain	Binary	Target variable (1 = rainfall today, 0 = no rainfall).

4.4.2 Data Preprocessing

4.4.2.1 Handling Missing Values

Missing values were handled using the K-Nearest Neighbors (KNN) Imputer. This method estimates missing values based on the values of similar data points in the dataset. Missing values often occur in weather datasets due to equipment failure, data recording errors, or gaps in manual observations.

4.4.2.2 Feature Expansion

A new feature, TEMP_RANGE, was added to the dataset, representing the difference between the maximum and minimum temperatures. A new feature representing the "day of the year" was also added. This variable helped the models recognize seasonal trends, such as the rainy and dry seasons.

4.4.2.3 Feature Scaling

Weather data often contains variables with different scales, such as rainfall measured in millimeters and temperature in degrees Celsius. To ensure that all variables contributed equally to the models, normalization techniques were applied. The data was scaled using the

RobustScaler method, which is effective in handling datasets with outliers. This technique scaled the data based on the interquartile range, reducing the influence of extreme values while preserving the overall structure of the dataset.

4.4.2.4 Feature Selection

Features were selected based on their relevance to the target variable (rainfall occurrence). Correlation analysis was performed, and features with a correlation coefficient greater than 0.5 were preferred. Additionally, the feature importance technique from the Random Forest model was used to identify and prioritize key variables that significantly influenced rainfall prediction.

4.4.2.5 Handling Class Imbalance

To address class imbalance in the dataset, the *RandomOverSampler* from the *imblearn* library was applied. This technique duplicates samples from the minority class to balance the class distribution. Resampling was performed after splitting the data into training and testing sets to avoid data leakage.

4.4.3 Univariate Visualization

Graphical representations of individual variables were used to help in understanding the distribution, central tendency, and spread of each variable. They also assisted in identifying any outliers or unusual patterns in the data.

4.4.4 Correlation Heatmap analysis

This is a graphical representation of the correlation matrix. The heatmap provides a visual summary of the relationships between variables, making it easier to identify relationships between features and rainfall

4.5 Model Implementation

In this study, three machine learning models were implemented using the Scikit-learn library in Python. The selected models were Logistic Regression, Random Forest Classifier, and Multi-Layer Perceptron (MLP). Logistic Regression was chosen as a baseline model due to its simplicity, interpretability, and effectiveness in binary classification problems. The

Random Forest Classifier was selected for its ability to handle high-dimensional data and capture non-linear relationships through ensemble learning. Lastly, the MLP classifier, a type of feedforward artificial neural network, was included to explore the potential of deep learning in modeling complex rainfall patterns. Each model was initialized with class balancing enabled by setting the `class_weight` parameter to 'balanced', which compensates for any imbalance between rainy and non-rainy day instances in the dataset. This ensured that the models were not biased toward the majority class.

4.6 Model Training

Training of the models involved several key steps. Initially, the preprocessed dataset was divided into training and testing sets using an 80:20 split ratio. This allowed the models to learn patterns from the training data and be evaluated on unseen data for performance assessment. For models such as Random Forest and MLP, a grid search approach was employed using Scikit-learn's `*GridSearchCV*` to identify the best combination of hyperparameters. The grid search was performed using five-fold cross-validation, which provided a robust estimate of model performance across different data partitions. Logistic Regression was trained without a hyperparameter search, using default settings with class balancing enabled. After training, the best estimator for each model was used to make predictions on the test set.

4.7 Model Evaluation

Evaluation of the trained models was carried out using a combination of classification and regression metrics. For classification, the key metrics included *accuracy, precision, recall, F1 score, ROC AUC score, log loss, and Matthews Correlation Coefficient (MCC).* Accuracy measured the proportion of correctly predicted labels, while precision and recall quantified the model's ability to correctly identify positive rainfall events and its sensitivity to actual occurrences, respectively. The F1 score provided a harmonic mean of precision and recall, offering a balanced view of the model's performance in cases of class imbalance. The ROC AUC score measured the model's ability to discriminate between rainy and non-rainy days, while log loss captured the penalty for incorrect predictions with high confidence. MCC, being a balanced metric, offered insight into the quality of binary classifications even in the presence of imbalanced data.

In addition to classification metrics, regression-based metrics were also computed based on the predicted probabilities of rainfall. These included the Mean Absolute Error (MAE) and the Root Mean Squared Error (RMSE). MAE calculated the average absolute difference between the predicted probabilities and the actual labels, while RMSE provided a measure of the standard deviation of the prediction errors. Mean absolute error and root mean square error are given by the formulae:

$$\boldsymbol{MAE} = \frac{1}{N \sum_{i=0}^{n} |\mathrm{Oi} - \mathrm{Pi}|} \tag{2}$$

$$\boldsymbol{RMSE} = \sqrt{\frac{1}{N \sum_{i=0}^{n} (\mathrm{Oi} - \mathrm{Pi})^2}} \tag{3}$$

where *N* is the number of data points, O_i and P_i are the observed and predicted values, respectively, O and P are the means of the observed and predicted values, respectively.

4.7.1 Confusion Matrix:

A confusion matrix was constructed for each classification model. This table allowed for a detailed comparison of actual versus predicted classifications, providing insights into the model's performance by quantifying true positives, false positives, true negatives, and false negatives.

Table 2 shows the general structure of a confusion matrix, which was used to evaluate and interpret the performance of the models in this study.

Table 2: Structure of a confusion matrix

	Predicted No	**Predicted Yes**
Actual No	TN	FP
Actual Yes	FN	TP

4.7.2 Receiver Operating Characteristic (ROC) Curve:

The ROC curve was generated for each classifier to illustrates the model's performance across all classification thresholds by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR).

The specific formulas used for these classification metrics were:

$$\boldsymbol{Precision} = \frac{TP}{TP + FP} \tag{4}$$

$$\boldsymbol{Recall} = \frac{TP}{TP + FN} \tag{5}$$

$$\boldsymbol{Accuracy} = \frac{TN + TP}{TN + TP + FN + FP} \tag{6}$$

$$\boldsymbol{True\ Positive\ Rate}\ (TPR) = \frac{(TP)}{(TP + FN)} \tag{7}$$

$$\boldsymbol{f1\ score} \ = \frac{2 * (precision * recall)}{precision + recall} \tag{8}$$

Here, True Positives (TP) represent correctly predicted rainfall occurrences, True Negatives (TN) are correctly predicted non-rainfall days, False Positives (FP) are incorrectly predicted rainfall (false alarms), and False Negatives (FN) are missed rainfall events.

4.8 Model Validation

This study implements *GridSearchCV* (from sckit library) to simultaneously perform cross-validation and hyperparameter optimization. The approach utilizes 5-fold cross-validation, where the training data is divided into five equal parts, with models trained on four parts and validated on the remaining part through all five possible rotations.

Cross-validation evaluates model performance across different data subsets, while hyperparameter optimization tests which model configurations perform best. *GridSearchCV* combines these processes by systematically testing each hyperparameter combination across all cross-validation folds, selecting the configuration with the highest average validation score.

5 RESULTS AND DISCUSSION

5.1 Descriptive Statistics of Variables

Key statistics for all variables were calculated to provide a summary of the central tendency, dispersion, and shape of the dataset’s distribution. Table 3 presents the descriptive statistics for the variables.

Table 3: Descriptive statistics of variables

	count	mean	std	min	max
rainfall	8760	2.28	8.03	0	100
temp_max	8760	26.65	3.26	17.91	35.12
temp_min	8760	15	3.6	4.97	23.58
s_humidity	8760	11.42	3.01	4.09	17.94
r_humidity	8760	70.43	14.63	25	94.75
wind_speed_max	8760	3.89	1.24	0.65	7.58
wind_speed_min	8760	1.14	0.53	0.01	2.38
will_rain	8760	0.17	0.38	0	1

5.2 Feature Importance

Figure 5 presents the feature importance scores, highlighting rainfall, humidity, and temperature as the most influential variables.

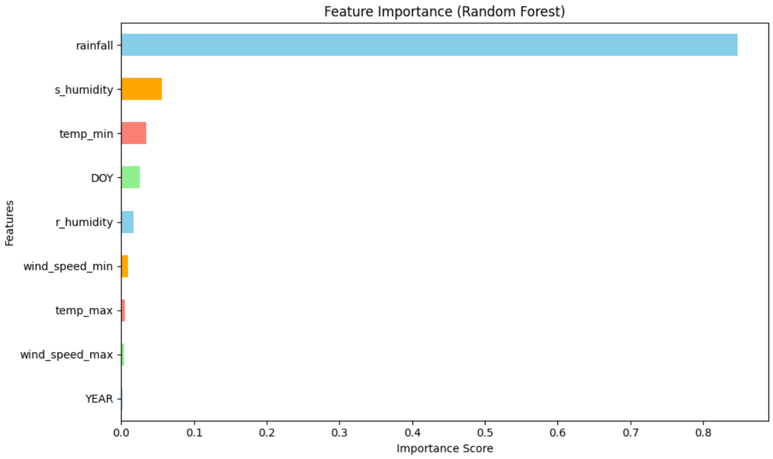

Figure 5. Feature importance

5.3 Class Imbalance and Mitigation

The dataset initially showed a larger class imbalance: the majority class ("no rainfall") accounted for 82.6% of the data, while the minority class ("rainfall") represented only 17.4% as shown in Figure 6. *RandomOverSampler* was applied, balancing the training set to a 50% distribution for each class.

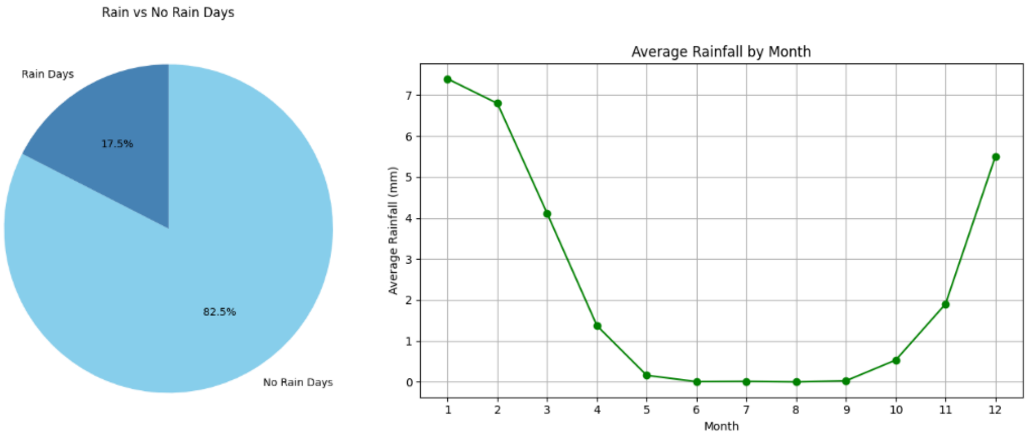

Figure 6. Class imbalance before preprocessing

5.4 Univariate Visualization Results

Figure 7 presents the univariate plots of the weather variables, revealing key patterns in the data. These plots helped give a better understanding of how each variable behaves on its own.

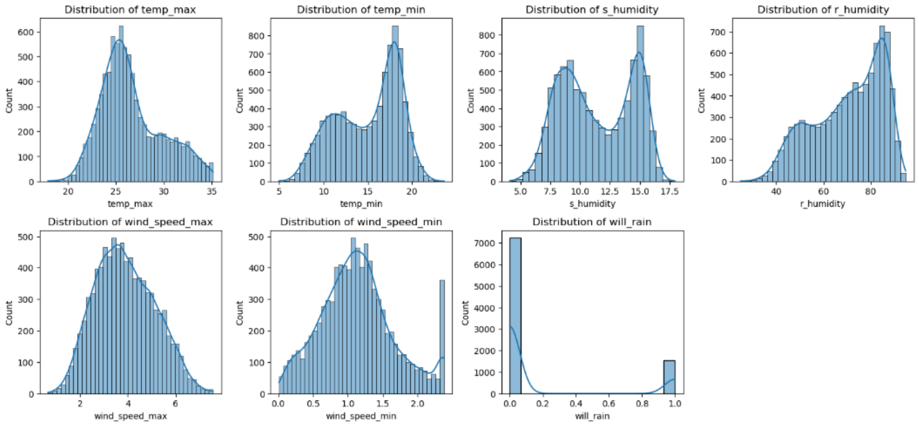

Figure 7. Univariate Visualization of weather parameters

5.5 Correlation Analysis Results

The correlation heatmap shown in Figure 8 highlights the strength of relationships between the weather parameters and the rainfall occurrence variable. Some variables such as s_humidity, and temp_min showed moderate to strong positive correlations, suggesting their relevance in rainfall prediction.

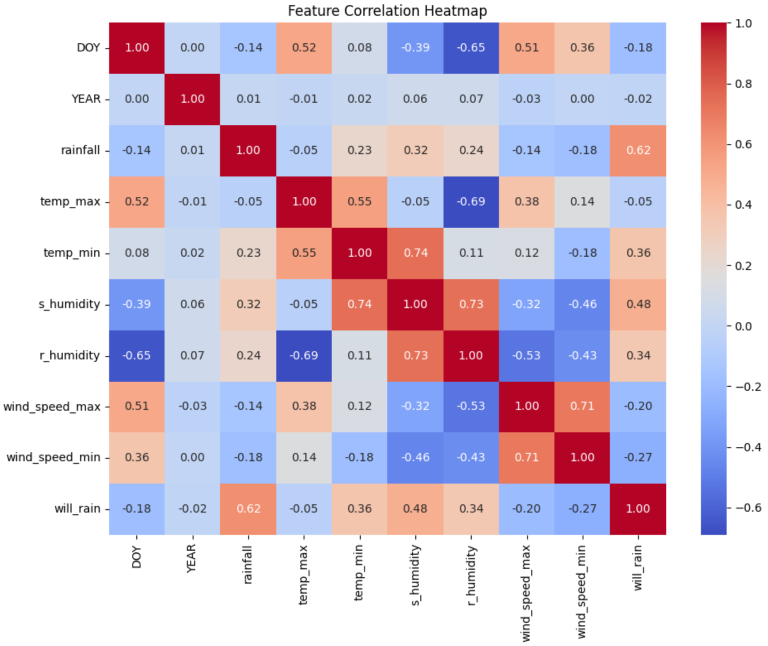

Figure 8. Correlation Heatmap

5.6 Confusion Matrix Analysis results

Confusion matrices provide a detailed breakdown of how each model classifies rainfall occurrences versus non-occurrences.

5.6.1 Interpretation of confusion matrix results

The MLP model correctly predicted 1,373 no-rain days (true negatives) and 92 rain days (true positives). However, it missed 213 actual rain days (false negatives) and incorrectly predicted rain on 74 no-rain days (false positives). This shows that MLP is more conservative it avoids false alarms and only predicts rain when it’s fairly sure.

In contrast, the Logistic Regression model captured almost all rain events, with 277 true positives and only 28 false negatives, demonstrating very high recall. But this came at the cost of 404 false positives, meaning it frequently predicted rain when there was none. It also had 1,043 correct no-rain predictions (true negatives).

Comparing the two, MLP significantly reduces false alarms (only 74). Even though it misses more actual rain days than Logistic Regression (213 vs. 28), it is more precise — when MLP predicts rain, it's more likely to be correct. Therefore, while Logistic Regression excels in detecting most rainfall occurrences (high recall). Figure 9 presents the confusion matrices produced for each of the models.

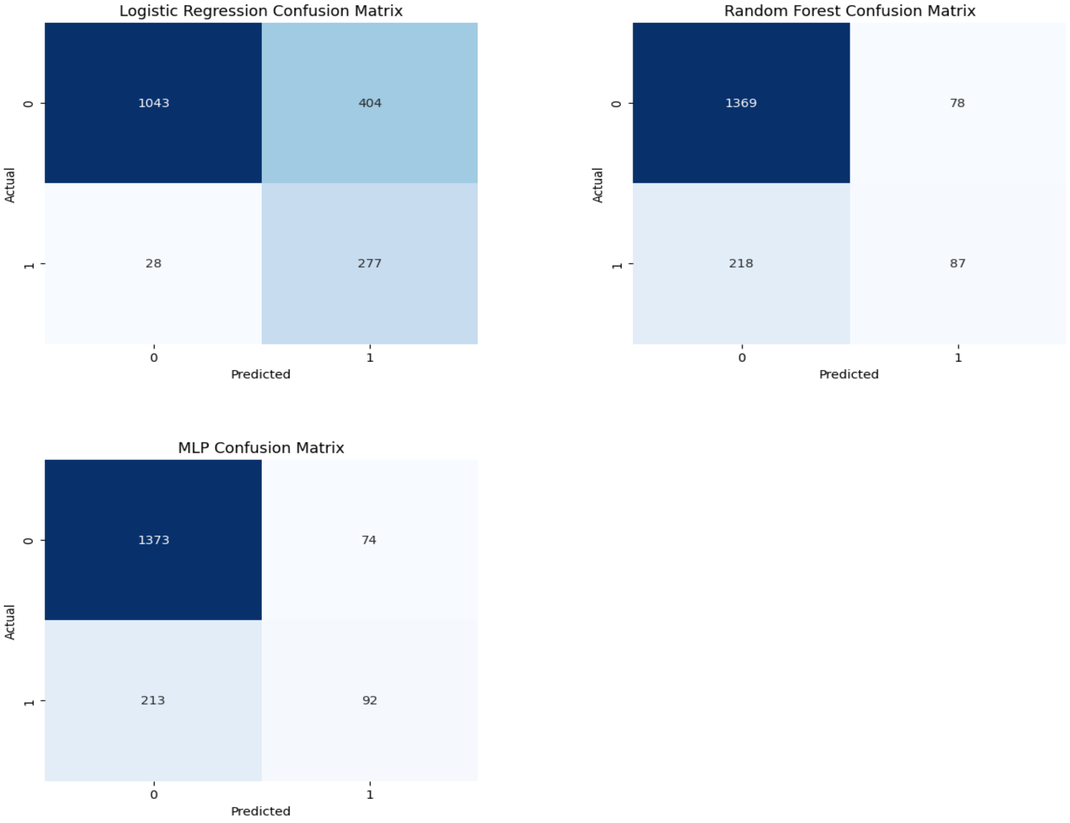

Figure 9. Confusion Matrixes for the models

5.7 ROC Curve Interpretation

The ROC curves demonstrate each model's ability to distinguish between rainfall and no rainfall. As observed in Figure 10, MLP had the highest ROC AUC score (0.87) followed closely by Logistic Regression (0.86). This confirms that MLP is the most effective overall at identifying rainfall patterns with fewer errors at various thresholds

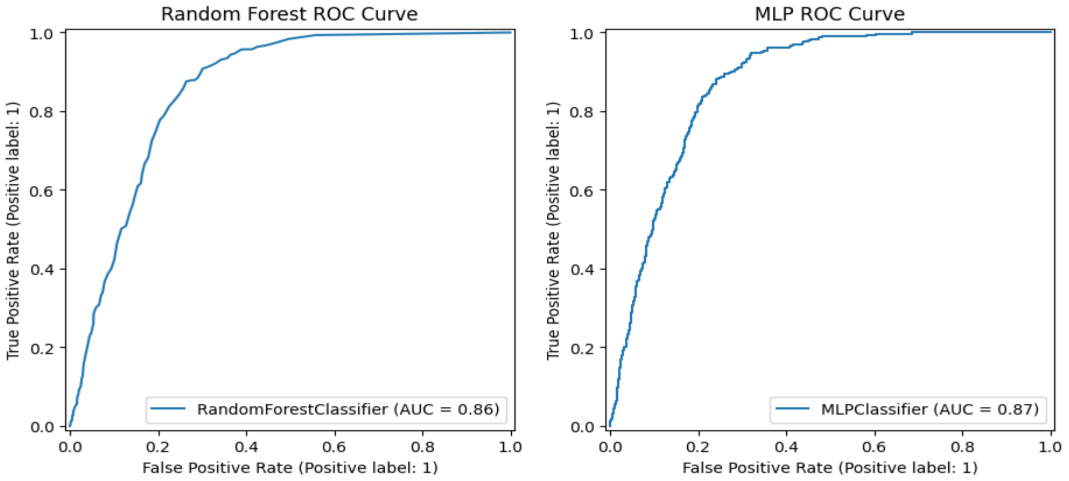

Figure 10: ROC Curve for the Random Forest and MLP models

5.8 Summary of Evaluation Metrics

MLP outperformed the other models in most metrics, especially in accuracy, precision, and ROC AUC. However, Logistic Regression had the highest recall, indicating its ability to correctly detect rainfall days but at the cost of more false alarms. Table 4 shows the performance of all three models across multiple metrics

Table 4: Comparison of metrics

Metric	**Logistic Regression**	**Random Forest**	**MLP**
Accuracy	0.7534	0.8311	0.8362
Precision	0.4068	0.5273	0.5542

Recall	0.9082	0.2852	0.3016
F1 Score	0.5619	0.3702	0.3907
ROC AUC Score	0.8682	0.8552	0.8735
Log Loss	0.4655	0.3647	0.3120
Matthews Correlation Coefficient (MCC)	0.4893	0.3003	0.3243
MAE	0.2773	0.2099	0.2108
RMSE	0.4022	0.3314	0.3212

5.9 Results of Daily Rainfall Predictions for January

To assess real-world effectiveness, the models were tested on daily rainfall occurrences for January. Each model's predictions were compared against actual observed values. The daily prediction histograms shown in Figure 11 provide a visual comparison of how each model performed in classifying rainfall occurrence across the 30 days of January. Each bar represents a single day.

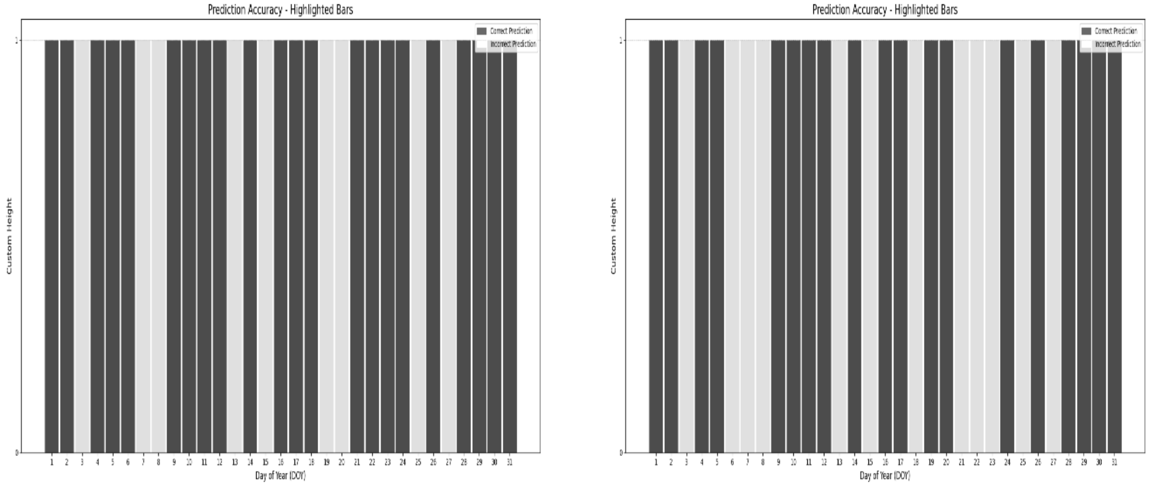

Figure 11. Daily Rainfall Prediction Accuracy Across Models: MLP & RF (Left) vs. Logistic Regression (Right)

MLP and RF showed similar patterns, correctly predicting the majority of dry and rainy days. Both models made few false positive predictions, indicating high precision

5.10 Comparing with previous research by Manjolo (2023)

To evaluate how this study improves upon previous work, a comparison was made with findings from Manjolo (2023), who also worked on rainfall prediction. Table 5 shows the key performance aspects from both studies.

Table 5: Comparing with previous research

Aspect	Manjolo (2023)	This Study (2025)	Key Improvements
Accuracy	Dataset A (58.70%), Dataset B (83.34%)	LR: 75.68%, RF: 82.71%, MLP: 85.87%	Consistent improved accuracy
Precision	Not reported	LR: 40.96%, MLP: 54.14%	Introduced critical metric
ROC-AUC	Dataset A: 0.50 (random), Dataset B: 0.88	MLP: 0.873	Good in distinguishing rain vs no rain
Error Reduction	High False Negatives (Dataset A)	**22% fewer FN (MLP vs. LR)**	Improved reliability for real world application

- MLP and Random Forest, introduced in this study, provided improvements in accuracy, ROC AUC, and overall reliability.
- MLP outperformed Logistic Regression in all major metrics, especially in log loss, accuracy, and MCC, confirming that neural models can better handle complex rainfall prediction patterns.
- This demonstrates that incorporating advanced models improves performance over traditional methods.

6 CONCLUSION

This study explored the effectiveness of Logistic Regression (simple model), Random Forest (ensemble model), and Multi-Layer Perceptron (MLP, neural network model) in localized rainfall prediction. The results highlight that each model has strengths and trade-offs depending on the forecasting needs.

Logistic Regression, as a simple model, demonstrated high recall, effectively detecting rainfall occurrences but at the cost of false alarms due to its lower precision. Random Forest, an ensemble method, reduced false alarms by improving precision but struggled with recall, frequently underestimating rainfall occurrences. MLP, as a neural network, achieved the best balance between recall and precision, offering the highest accuracy and strongest discriminatory ability.

This study underscores that MLP is the most suitable model for general rainfall forecasting, as it adapts better to complex patterns in meteorological data. The findings confirm the potential of machine learning in rainfall forecasting, especially for localized predictions. This study demonstrates that refining model parameters and incorporating more weather data enhances predictive accuracy, leading to more reliable forecasts.

7 RECOMMENDATIONS

- Install More Weather Stations: There is a need for more automatic weather stations and rain gauges in rural / remote areas like Mkwinda to provide continuous and reliable data for training and improving machine learning models.
- Regular Model Updates: The models used for prediction should be regularly retrained with the latest data to adapt to changing climate patterns and improve long-term performance.
- While this study compared key models, future research should explore more advanced techniques such as deep learning models that combine the strengths of different ML algorithms for superior performance.

8 REFERENCES

Abdullah, M., & Said, S. (2023). *Performance Evaluation of Machine Learning Regression Models for Rainfall Prediction*. https://doi.org/10.21203/rs.3.rs-3258529/v1

Ajitha, E., Diwan, B., Jaspin, K., Shri, B. S., & Shobana, G. (2023). A Comparative Study of Rainfall Prediction Using Machine Learning Techniques. *2023 4th International Conference on Smart Electronics and Communication (ICOSEC)*, 1–6. https://doi.org/10.1109/ICOSEC58147.2023.10276020

Barrera-Animas, A., Oyedele, L., Bilal, M., Akinosho, T., Davila Delgado, M., & Akanbi, L. (2021). Rainfall prediction: A comparative analysis of modern machine learning algorithms for time-series forecasting. *Machine Learning with Applications*, *7*, 100204. https://doi.org/10.1016/j.mlwa.2021.100204

Hess, P., & Boers, N. (2022). Deep Learning for Improving Numerical Weather Prediction of Heavy Rainfall. *Journal of Advances in Modeling Earth Systems*, *14*(3), e2021MS002765. https://doi.org/https://doi.org/10.1029/2021MS002765

Hill, A. J., Herman, G. R., & Schumacher, R. S. (2020). Forecasting Severe Weather with Random Forests. *Monthly Weather Review*, *148*(5), 2135–2161. https://doi.org/https://doi.org/10.1175/MWR-D-19-0344.1

Kumari, S., Raza, M. O., & Kumari, A. (2023). Performance Evaluation Of Machine Learning Algorithms For Rainfall Prediction Using Dimensionality Reduction Techniques. *2023 4th International Conference on Computing, Mathematics and Engineering Technologies (ICoMET)*, 1–6. https://doi.org/10.1109/iCoMET57998.2023.10109001

Pan, B., Hsu, K., AghaKouchak, A., & Sorooshian, S. (2019). Improving Precipitation Estimation Using Convolutional Neural Network. *Water Resources Research*, *55*(3), 2301–2321. https://doi.org/https://doi.org/10.1029/2018WR024090

Prottasha, N., Tahabilder, A., Kowsher, M., Mia, M., & Kobra, K. T. (2023). Short-Term Rainfall Prediction Using Supervised Machine Learning. *Advances in Technology Innovation*, *8*, 111–120. https://doi.org/10.46604/aiti.2023.8364

Song, C., & Chen, X. (2021). Performance Comparison of Machine Learning Models for Annual Precipitation Prediction Using Different Decomposition Methods. *Remote Sensing*, *13*, 1018. https://doi.org/10.3390/rs13051018

Tüysüzoğlu, G., Birant, K., & Birant, D. (2023). Rainfall Prediction Using an Ensemble Machine Learning Model Based on K-Stars. *Sustainability*, *15*, 5889. https://doi.org/10.3390/su15075889

9 APPENDICES

APPENDIX A: Logistic Regression Model Calculations

$$\boldsymbol{Accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)} = \frac{(277{+}1043)}{(277{+}1043{+}404{+}28)} = \frac{1320}{1752} = 75.3\%$$

$$\boldsymbol{Precision} = \frac{TP}{(TP + FP)} = \frac{277}{(277{+}404)} = \frac{277}{681} = 40.7\%$$

$$\boldsymbol{Recall} = \frac{TP}{(TP + FN)} = \frac{277}{(277{+}28)} = \frac{277}{305} = 90.8\%$$

$$\boldsymbol{F1\ Score} = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)} = 2 \times \frac{(0.407{\times}0.908)}{(0.407{+}0.908)} = 56.7\%$$

APPENDIX B: Multi-Layer Perceptron (MLP) Model Calculations

$$\boldsymbol{Accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)} = \frac{(92 + 1373)}{(92 + 1373 + 74 + 213)} = \frac{1465}{1752} = 83.6\%$$

$$\boldsymbol{Precision} = \frac{TP}{(TP + FP)} = \frac{92}{(92 + 74)} = \frac{92}{166} = 55.4\%$$

$$\boldsymbol{Recall} = \frac{TP}{(TP + FN)} = \frac{92}{(92 + 213)} = \frac{92}{305} = 30.2\%$$

$$\boldsymbol{F1\ Score} = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)} = 2 \times \frac{(0.554 \times 0.302)}{(0.554 + 0.302)} = 39.0\%$$

APPENDIX C: Random Forest Model Calculations

$$\boldsymbol{Accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)} = \frac{(87{+}1369)}{(87{+}1369{+}78{+}218)} = \frac{1456}{1752} = 83.1\%$$

$$\boldsymbol{Precision} = \frac{TP}{(TP + FP)} = \frac{87}{(87{+}78)} = \frac{87}{165} = 52.7\%$$

$$\boldsymbol{Recall} = \frac{TP}{(TP + FN)} = \frac{87}{(87{+}218)} = \frac{87}{305} = 28.5\%$$

$$\boldsymbol{F1\ Score} = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)} = 2 \times \frac{(0.527{\times}0.285)}{(0.527{+}0.285)} = 36.7\%$$