IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 17 April 2024, accepted 27 April 2024, date of publication 3 May 2024, date of current version 10 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3396695

RESEARCH ARTICLE

Analysis of Machine Learning Algorithms for Prediction of Short-Term Rainfall Amounts Using Uganda's Lake Victoria Basin Weather Dataset

TUMUSIIME ANDREW GAHWERA[1], ODONGO STEVEN EYOBU[2], AND MUGUME ISAAC[3]
[1]Department of Information Systems, School of Computing and Informatics Technology, Makerere University, Kampala, Uganda
[2]Department of Networks, School of Computing and Informatics Technology, Makerere University, Kampala, Uganda
[3]College of Agricultural and Environmental Sciences, Makerere University, Kampala, Uganda

Corresponding author: Odongo Steven Eyobu (odongo.eyobu@mak.ac.ug)

This work was supported by Makerere University, Research and Innovation Fund (RIF) under the Government of Uganda.

ABSTRACT As a result of climate change, the difficulty in the prediction of short-term rainfall amounts has become a necessary area of research. The existing numerical weather prediction models have limitations in precipitation forecasting especially due to high computation requirements and are prone to errors. Precipitation amount prediction is challenging as it requires knowledge on a variety of environmental phenomena, such as temperature, humidity, wind direction, and more over a long period of time. In this study, we first of all present our Lake Victoria Basin weather dataset and then use it to conduct a rigorous analysis of machine learning algorithms to do short-term rainfall prediction. The rigorous analysis includes algorithm optimizations to improve prediction performance. In particular, we validate our weather dataset using various machine learning regression models which include Random Forest regression, Support Vector regression, Neural Network regression, Least Absolute Shrinkage and Selection Operator regression, Gradient boosting regression, and Extreme Gradient boosting regression. The performance of the models was evaluated using Mean Absolute Error (MAE), and Root Mean Square Error (RMSE). The findings demonstrate that, in comparison to other algorithms, Extreme Gradient Boost regression has the lowest MAE values of 0.006, 0.018, 0.005 for Lake Victoria basin weather data in Uganda, Kenya, and Tanzania respectively.

INDEX TERMS Precipitation amount, weather prediction, data-driven approaches, short-term forecasting.

I. INTRODUCTION

Weather forecasts are extremely important as many industries such as agriculture, shipping, engineering, construction, natural disasters, aviation, and defence rely heavily on weather dynamics for successful operations [1]. For example, while deciding whether to plant, weed, or spray plants or animals, a farmer has to know if it will rain or not. When deciding whether to take off, pilots must carefully consider the day's weather forecast. The students and those in other sectors, such as industries, need daily weather forecasts to make timely decisions. Precipitation is one of the important measurement problems for many weather forecasting applications. However, due to its high spatial and temporal variability, precipitation is one of the most difficult weather variables to predict. The chaotic nature of the atmosphere makes accurate weather forecasting, particularly precipitation amounts extremely difficult [2].

For short-, medium- and long-term climate forecasting, numerical weather prediction (NWP) models are used. The Weather Research and Forecasting (WRF) model is the recommended numerical weather forecast model for Uganda, with the ability to estimate daily, weekly and monthly rainfall [3]. According to [4], this model is the most

The associate editor coordinating the review of this manuscript and approving it for publication was Chao Tong.

up-to-date and most widely used mesoscale model in the world, by both the academic and operational forecasting communities.

WRF model results can be automated and checked with advanced tools such as atmospheric assessment models. This simplifies configuration for operational use. It is important to note that NWP models require large datasets [1], which in turn requires expensive physical hardware and significant computational power [5].

Currently, machine learning (ML) methods are being explored in many other fields, with great success in image processing [6], medical diagnosis [7], pattern recognition [8] and time series prediction [9], weather forecasting in space and time [10] and other fields.

In atmospheric science, where huge heterogeneous databases exist and are accessible, data-driven techniques are used. In [5], artificial intelligence (AI) is used by the US National Oceanic and Atmospheric Administration (NOAA) to improve the efficiency, accuracy, and synchronization of NOAA services. One application of machine learning techniques used by Météo-France is storm forecasting. The UK's European Centre for Medium-Range Weather Forecasts (ECMWF) has used machine learning to improve its numerical forecasts. In this work, we present the prospect of advanced ML methods in weather forecasting, which can replace numerical solutions to differential equations such as [11].

One of the benefits of using machine learning for weather and climate forecasting is reduced computing power [1]. Several researchers have used meteorological data from many countries to perform experiments to improve daily, monthly, and yearly precipitation prediction. Researchers used big data analysis techniques [12], machine learning [13] and data mining techniques [14] to increase the accuracy of daily rainfall forecasts, monthly and yearly. Thus, researchers [7], [15], [16], confirm that machine learning algorithms outperform conventional deterministic methods in forecasting weather and precipitation.

The authors in [17] predict daily rainfall in Ethiopia using regression and classification algorithms such as Decision Trees (DT), Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Multi-layer Perceptron (MLP) and Long-Short-Term Memory (LSTM). Their results show that LSTM significantly outperforms regression algorithms in predicting daily precipitation. Researchers in [18] used various deep learning and machine learning models to calculate rainfall totals, predict rainfall frequency, and evaluate the performance of regression and classification models. Their approach outperforms other advanced techniques for predicting precipitation.

The gradient boost regression (GBR) model is also effectively used in [19]'s work to estimate daily evapotranspiration and irrigation planning. In [20], authors applied statistical and machine learning methods; multiple linear regression (MLR), support vector regression (SVR) and least absolute shrinkage and selection operator (LASSO) for precipitation forecasting. Notably, statistical methods were surpassed by machine learning algorithms.

In this study, we experimented the effectiveness of different machine learning methods in forecasting short-term hourly rainfall amounts around the Lake Victoria basin in Uganda using our published dataset [21] as a basis.

The study trained and tested six machine learning regression algorithms namely; Random Forest (RF), Support vector regression (SVR), Neural network regression (NNR), Gradient boost regression (GBR), LASSO regression, and Extreme gradient boosting regression (XGBoost) using the Lake Victoria basin weather dataset with the purpose of finding out the best performing regression model. From the experiments XGBoost was the best performing model among the models tested specifically in terms of forecasting hourly precipitation amounts.

In addition, the XGBoost model was tested for generalizability in areas with comparable weather patterns using datasets from the Lake Victoria basin in Kenya and Tanzania. XGBoost optimization model was still the best performing model in these two adjacent nations.

The research contributions of this study are as follows:

(i) A review of machine learning models used for prediction of weather conditions.
(ii) An analysis of different machine learning models for prediction of short-term rainfall amounts over the Lake Victoria basin.
(iv) A statistical comparison of various machine learning algorithms used for precipitation amount prediction.
(iv) A highlight of potential future study areas in the application of machine learning to improve precipitation amount predictions.

The rest of this paper is organized as follows. Section II presents an overview of relevant works done by various researchers. Section III presents the technical preliminaries on the selected machine learning models. Section IV describes the analysis of ML models for short-term rainfall amounts as well as the dataset used for the study. Section V results and discussions. Section VI conclusion and future work.

II. RELATED WORKS ON WEATHER PREDICTION

In this section, we present comprehensive literature on machine learning models and identify the utilized models and their applications. The applied models are discussed throughout this study. Table 1 summarizes the related studies done using regression-based forecasting methods. The table shows the contribution of each study, the techniques, and the application domain. Table 2 evaluates the performance of different metrics reviewed with the aim of identifying gaps from related studies.

Over the past years, there have been a precedented increase in the use of artificial intelligence methods in weather and climate research [22]. This has been made possible

TABLE 1. A summary of related works in short-term weather forecasting.

Reference	Contribution	Techniques	Application Domain
Liyew and Melese, [10]	developed a machine learning method to forecast Ethiopia's daily precipitation totals.	MLR, RF and XGBoost	Rainfall prediction
Balamurugan and Manojkumar, [13]	proposed a machine learning-based approach to anticipate short-term rainfall	Logistic regression, DT, RF and statistics	Rainfall prediction
Endalie et al. [17]	developed a deep learning model to forecast daily precipitation	DT, SVM, KNN and MLP	Precipitation forecasting
Kanani et al. [18]	employed various machine learning and deep learning models to: (a) forecast the frequency of rainfall; (b) estimate the total amount of rainfall; and (c) evaluate the outcomes of the various models for regression and classification	Polynomial Regressor, MLR, XGBoost, RF and LSTM	Rainfall prediction
Mohammed et al. [20]	applied machine learning regression approaches to predict precipitation	MLR, SVR and statistical techniques	Precipitation prediction
Karna et al. [25]	investigated the use of linear regression machine learning algorithm for time series data	LR	Weather forecasting
Ponraj and Vigneswaran, [19]	proposed a gradient boost regression model to forecast daily evapotranspiration and plan irrigation	MLR, RF and GBR	Evapotranspiration prediction
Zheng and Wu, [29]	proposed a new extreme gradient boosting model with weather similarity analysis and feature engineering for short-term wind power forecasting	CART, RF, SVR and single XGBoost	Wind power forecasting
Deng et al. [30]	proposes the Bagging-XGBoost algorithm based extreme weather identification and short-term load forecasting model, which can warn the time period and detailed value of peak load in advance	Bagging-XGBoost	Load forecasting
Ma et al. [31]	developed and evaluated a Xgboost model for the prediction of outdoor air temperature and humidity using acquired data from Shenzhen	XGBoost	Temperature Forecasting
Oleiwi et al. [58]	searched on regional precipitation pattern modeling using three intelligent predictive models incorporating ANN, SVM, and RF methods	ANN, SVM and RF	Precipitation forecasting
Kareem et al. [63]	identified various ANN architectures for weather forecasting	CNN, LSTM and BPNN	Weather forecasting
Krammer et al. [64]	presented machine learning models to estimate rainfall in the geographical region of central Europe, in Slovakia	NN, RF and SVM	Rainfall prediction
Lawal et al. [65]	evaluated the different machine learning models for rainfall prediction using Nyando in Kenya as a case study	LSTM, XGBoost, RF and SVR	Rainfall prediction
Ganapathy et al. [66]	applied six different ML algorithms using a rainfall dataset of the Vellore region, of Tamil Nadu, India for the years 2021 and 2022 to predict rainfall	ARIMA, LSTM, SVR, LR, XGBoost and Holt-Winters' Exponential Smoothing	Rainfall prediction

through low-cost data gathering devices including sensors, satellite cameras, rain gauges, and more that are deployed globally to record real-time data. Second, the current ease of implementing machine learning techniques is a result of the creative use of computer graphical processing units (GPUs), which are faster with improved computational power than standard central processing units (CPUs) [7].

Machine learning, then, is a branch of artificial intelligence that uses statistical models and algorithms trained to identify patterns in big training datasets to generate accurate predictions. Machine learning models are categorised as regression or classification problems [23]. In this study, we are addressing a regression problem with the aim of predicting hourly precipitation amounts over the Lake Victoria basin in Uganda.

Several research investigations have applied machine learning regression models in weather forecasting. For example, machine learning literacy algorithm was used in [24] to forecast the weather. The purpose of the study was to produce accurate weather prediction. The study made use of the following variables: date, temperature on that specific date, wind speed and direction, maximum and minimum temperature, and weather. Rainfall was predicted using the machine learning linear regression model.

The study in [25] investigated the use of linear regression (LR) machine learning algorithm for time series data. The primary goal of the research was to develop long-term temperature prediction for Pokhara's numerous weather stations. The Department of Hydrology and Meteorology (DHM) in Pokhara provided the dataset with which the machine learning models were trained. The analysis of maximum temperature prediction model based on regression offered superior accuracy.

Multiple linear regression model was suggested by the authors in [26] as a means of predicting Bangladesh's rainfall. Using data mining, the objective was to accurately forecast rainfall. Four meteorological factors; precipitation, cloud cover, average temperature, and vapor pressure were obtained from Rajshahi, Bangladesh. The results indicated multiple linear regression as the best prediction model.

The study in [17] created a deep learning model to forecast daily precipitation. The study's objective was to develop a rainfall prediction model for Jimma, an area in Ethiopia's southwest province of Oromia. The Long Short-Term Memory based prediction served as the foundation for the proposed model. The suggested model was evaluated against several machine learning regressions, including Decision Tree, Support Vector Machine, K-Nearest Neighbors, and Multi-layer Perceptron. Based on the findings, the suggested LSTM model performed better than the alternative regression techniques.

Kanani et al. [18] employed various machine learning and deep learning models to: (a) forecast the frequency of rainfall; (b) estimate the total amount of rainfall; and (c) evaluate the outcomes of the various models for regression and classification. Data from 49 Australian cities over a ten-year period make up the dataset utilized in this work to predict rainfall. It includes 23 features, such as location, temperature, evaporation, daylight, wind direction, and many more. For the classification task, the research used training classifiers like XGBoost, RF, Kernel SVM, and LSTM. For the regression task, models like Multiple Linear Regressor, XGBoost, Polynomial Regressor, Random Forest Regressor, and LSTM were used. The results demonstrated that the suggested methodology surpasses various state-of-the-art methods.

The work of Ponraj and Vigneswaran [19] used a gradient boost regression model to forecast daily evapotranspiration to aid irrigation. Evapotranspiration was projected using daily weather data for the lowest and maximum temperatures, relative humidity, solar radiation, soil temperature, and wind speed. The GBR, RF and MLR algorithms were used to train, validate, and test the datasets. The root mean square error, mean absolute error, and coefficient of determination were compared in order to assess the effectiveness of these models. In terms of evapotranspiration prediction, it was discovered that the pre-processed GBR model outperforms the other two models.

Machine learning regression approaches are suggested by authors in [20] to predict precipitation. The aim was to give a comparative analysis of the different machine learning algorithms used in the field of precipitation prediction easily accessible to non-experts. Multiple linear regression, Support vector regression, and statistical techniques were utilized in the study, and the different machine learning methodologies were compared and analysed. According to the data, SVR had the best results compared to MLR and Lasso regression.

Machine learning-based approach to anticipate short-term rainfall is also suggested in research [13]. From data acquired by the Indian Meteorological Department (IMD), the study took into account variables such as temperature, precipitation, atmospheric pressure, humidity and wind speed. The study's findings and the statistically derived numerical projections were contrasted. The results showed that machine learning techniques outperformed classical numerical techniques.

Liyew and Melese [10] developed a machine learning method to forecast Ethiopia's daily precipitation totals. The study's primary goal was to pinpoint the pertinent atmospheric elements that contribute to rainfall and utilize machine learning techniques to forecast the amount of rain that will fall each day. Relevant environmental variables were chosen using the Pearson correlation technique and fed into the machine learning model. The local meteorological office in Bahir Dar City, Ethiopia provided the dataset, which was used to assess how well three machine learning methods performed. Extreme Gradient Boost, Random Forest, and Multivariate Linear Regression were the models used. According to the results, the machine learning method extreme gradient boosting outperformed the other competing approaches. The researchers did, however, note that by combining sensor and meteorological datasets and applying big data analysis techniques, future research could significantly increase forecast accuracy.

Garg and Pandey [28] presented a machine learning-based rainfall prediction. Utilizing SVR, SVM, and KNN machine learning algorithms. The goal was to forecast the amount of rainfall for the upcoming year and compare the conclusions drawn from each technique. The dataset was acquired between 1951 and 2015 through the National Data Sharing and Accessibility Policy (NDSAP) of India. The results of the experiment demonstrated that SVM was the best of the three, and that the best approach to apply it is to add bias to the model in order to produce a range of highest and lowest projected values.

Gupta et al. [8] suggested employing machine learning to forecast rainfall. In order to create weather forecasting models that predict whether it will rain in major cities tomorrow based on the day's meteorological data, the study's objective was to implement rainfall prediction using machine learning techniques like RF, DT, logistic regression and Neural network (NN). The primary data source was Kaggle/Twitter. The experimental findings demonstrated that RF beat logistic regression, NN, and DT models.

Zheng and Wu [29] proposed a new Extreme gradient boosting model with weather similarity analysis and feature engineering for short-term wind power forecasting. The authors considered similarities among historical days weather using K-means clustering algorithm to divide the samples into several categories. The XGBoost model was used to predict for each category. The results of the proposed model are compared with the back propagation neural network (BPNN) and classification and regression tree (CART), RF, SVR and a single XGBoost model. XGBoost was found to be the best model for short-term wind speed forecasting.

Deng et al. [30] proposes the Bagging–XGBoost algorithm based extreme weather identification and short-term load forecasting model, which can warn the time period and detailed value of peak load in advance. Firstly, using XGBoost algorithm, the idea of Bagging is introduced to reduce the output variance and enhance the generalization

ability of the algorithm. The weather mutual information theory, the correlation between transformer load and various weather factors is analysed. The experimental findings demonstrate that the proposed model reduces the average Mean Absolute Percentage Error (MAPE) of peak load by 3% to 10%.

Ma et al. [31] developed and evaluated XGboost model for the prediction of outdoor air temperature and humidity using acquired data from Shenzhen. The purpose was to use XGboost to predict outdoor temperature and humidity in predictive horizon of 1-3 hours. The results show the excellent performance of XGboost in accurately predicting outdoor temperature and humidity by comparison between the measured and predicted outdoor air temperature and air humidity.

Oleiwi et al. [58] researched on regional precipitation pattern modeling using three intelligent predictive models incorporating Artificial Neural Network (ANN), SVM, and RF models. The study used monthly time scale precipitation data for semi-arid environment in Iraq. Regional data for twenty weather stations were used for predictive modeling. The area under study was split up into districts. Following that, weather data from each district was modeled, and then, each district's weather data was used to do cross-station modeling. The authors define cross-station as using data from one station to train models and a different station to perform predictions. The study revealed that cross-station modeling was very effective in predicting the spatial distribution of precipitation in watersheds with limited meteorological data.

The various ANN architectures for weather forecasting for a comparable geographic region were presented by Kareem et al. [63]. Convolutional Neural Network (CNN) and several ANN architectures, such as LSTM and BPNN algorithms, were assessed in the study. Every model that was chosen had its own certain qualities. Results showed that ANN algorithms outperformed CNN in terms of performance. This is because time series meteorological data can have non-linear temporal relationships captured by ANN algorithms.

Krammer et al. [64] presented machine learning models to estimate rainfall in the geographical region of central Europe, in Slovakia. To predict precipitation, the authors used several different types of models including regression trees, lazy methods, linear regression with kernels, SVM regression, among others. The study results demonstrated that RF and NN were the best models for this kind of scenario. Furthermore, RF was more robust using several input features including those with a low significance. However, the researchers suggested that future studies could improve prediction accuracy by using an ensemble of RF and NN models.

Lawal et al. [65] evaluated the different machine learning models for rainfall prediction using Nyando in Kenya as a case study. The study explored both univariate and multivariate models to improve prediction accuracy using the LSTM, XGBoost, RF, and SVR algorithms. The study revealed XGBoost multivariate model performed best in daily and monthly prediction, while LSTM models showed potential despite facing challenges in capturing long-term dependencies. However, the study found that univariate models could not identify complex relationships in the data compared to multivariate models.

Ganapathy et al. [66] study applied a rainfall dataset of the Vellore region, of Tamil Nadu, India. The duration of the dataset is from 2021 to 2022 to forecast rainfall using six different algorithms; the Auto-Regressive Integrated Moving Average (ARIMA) model, Holt-Winters' Exponential Smoothing, LSTM, SVR, XGBoost and Linear regression. In their work, feature engineering was used to create new features in order to eliminate all forms of auto-correlation from the data. Regression models were used to manipulate the dataset. In addition, the analysis employed MAE, RMSE, Root Relative Squared Error (RRSE), Relative Absolute Error (RAE) as evaluation metrics. From the results, XGBoost model was the best performing model on the test data. However, the authors suggest improving future rainfall forecasting using deep learning techniques to further enhance the precision of the forecasting models.

In table 2, the best performance metrics for the applied models were registered. These values are normal based on the dataset, techniques, complexity of the data, outliers, model selection, normalization, and data size. We observed that the MAE for all the models reviewed were lower than the other metrics in all studies. The R^2 values are used when the predicting features are linear as seen in [25]. We acknowledge that R^2 is a great performance metric only and only if there exist a linear relationship between the predictors and the target features. Nash-Sutcliffe Efficiency (NSE) is majorly used for hydrological water flow modeling [32] which may not be well utilized in weather prediction.

Therefore, we conclude that alternative metrics MSE, RMSE, MAE, and RMAE are considered mainly in analysis of weather data as performance metrics because they are not susceptible to outliers and can better reflect actual situation of the predicted error [29]. In our study, experiments were conducted on our dataset for Uganda's Lake Victoria basin [21]. The results revealed performance metrics MAE and RMSE as the best metrics. Consequently, these metrics were considered in evaluating the performance of our model.

III. TECHNICAL PRELIMINARIES

The applicable machine learning regression models employed in this investigation are briefly covered in this section. This category includes Random Forest regression, Support vector regression, Neural network regression, Least Absolute Shrinkage and Selection Operator regression, Gradient Boost regression and Extreme Gradient boost regression. These are discussed further in the following section.

A. RANDOM FOREST REGRESSION

We selected RF model because of its capability in handling complex datasets like weather data. It is also an ensemble

TABLE 2. Summary of performance metrics used in weather forecasting studies.

Reference	Technique	Perfomance metrics							
		MAE	RMSE	R^2	NRMSE	NSE	RMAE	MAPE	MSE
Liyew and Melese, [10]	XGBoost	3.58	7.85	x	x	x	x	x	x
Kanani et al. [18]	RF	0.117	x	0.76	x	x	x	x	x
Ponraj and Vigneswaran, [19]	GBR	0.13	0.20	0.984	x	x	x	x	x
Endalie et al. [17]	LSTM	0.0082	0.010	0.9972	0.018	0.81	x	0.4786	x
Ma et al. (2020)	XGBoost	x	< 0.81	> 0.73	x	x	x	0.4786	x
Karna et al. [25]	LR	1.78	3.10	x	x	x	x	x	x
Mohammed et al. [20]	SVR	4.35069	x	0.9958	x	x	x	x	x
Balamurugan and Manojkumar, [13]	DT	x	0.1126	x	x	x	x	x	x
Zheng and Wu, [29]	XGBoost	16.93	23.28	0.9958	x	x	4.11	x	542.11
Deng et al. [30]	Bagging-XGBoost	x	x	x	x	x	x	3-10 %	x
Lawal et al. [65]	XGBoost	0.042529	0.05654	x	x	x	x	x	x
Ganapathy et al. [66]	XGBoost	0.226879	0.461475	x	x	x	x	x	x
Proposed model	XGBoost	**0.00616**	**0.04439**	x	x	x	x	x	x

supervised learning model that predicts by selecting individual multiple decision tree regressors randomly [33]. The number of trees, input variables and node size all have an impact on the RF regression model's efficacy. This model splits data based on feature values to provide predictions. Following that, the mean response value of the samples in the current leaf node is used to calculate each sample's prediction. Albeit, RF has shown promise in forecasting high-dimensional data and handling missing values, RF, being an ensemble of decision trees, may struggle to capture temporal dependencies and trends present in time series weather data [34]. Weather patterns often involve sequential dependencies, and RF might not naturally capture these dynamics. Figure 1 depicts the RF architecture for classification and regression analysis.

B. NEURAL NETWORK REGRESSION

The dependent variable in our study, the precipitation rate per hour (Prec_rate), did not show any linear correlation with the remaining independent features. As a result, linear models were not considered for further investigation. Consequently, alternative approaches were required to learn non-linear complicated interactions, and Artificial Neural Networks (ANN) was one of these techniques since it uses the activation function in each layer [35]. Numerous researchers have considered Artificial Neural Networks in rainfall modeling [36]. Self-adaptive ANNs can address non-linearity issues in rainfall data without requiring any correlations with variables evaluated [37]. Precipitation forecasting, on the other hand, is a more difficult challenge since it is inhibited by temporal and spatial variations in regional rainfalls. Different forms of ANN such as Recurrent Neural Network (RNN) and Long Short-term Memory among others can be utilized to address these types of issues in rainfall modeling [37]. The general neural network architecture is shown in figure 2.

C. SUPPORT VECTOR REGRESSION

Support vector regression is a regression model designed for predicting a continuous target variable [39]. Support vectors are utilized for both regression and classification tasks. To address non-linearity in classification problems, kernel based SVM is used. In SVM, inputs are mapped into high-dimensional space and non-linear correlations between predictors and response variables are transformed into linear ones. Furthermore, SVM deals with categorization based on threshold and classes or labels [38]. However, SVM-based models are prone to overfitting during calibration and underfitting during validation [38]. Nonetheless, SVR was selected to address non-linear challenges in weather data modeling. To fit an SVR model, equation 1 must be employed to solve the optimization problem.

$$y = MIN\frac{1}{2}||W||^2 \text{ with the constraint of } |y_i - b_i x_i| \leq \epsilon \quad (1)$$

D. LASSO REGRESSION

Least Absolute Shrinkage and Selection Operator regression was utilized in this work because of its regularization abilities. The model's purpose is to reduce over fitting while improving model generalization. The goal is to identify a line (hyperplane) that best matches the data and limits the model's complexity by reducing the magnitude of the coefficients. Lasso is also employed in feature selection to find which predictors have the highest correlation with the outcome variable [40]. However, this model may have unbiased coefficients due to the Lasso regularization (L1 penalty), which artificially lowers the coefficients to zero. Consequently, the magnitude of the link between features and outcomes is not satisfied. The objective function for lasso regression is given in equation 2.

$$\lambda * (|\beta_1| + |\beta_2| + \ldots + |\beta_p|) \quad (2)$$

E. GRADIENT BOOSTING REGRESSION

Gradient boosting regression is an ensemble method built on the idea of decision trees [27]. The decision tree has a tree-like structure, with branches that branch out to the leaves depending on numerous parameters. The anticipated result is the goal leaf. The goal leaf (leaf node) is where the final predictions are assigned. The ensemble approach is used in decision trees to address the problem of overfitting the test data. So, instead of using a single decision tree, you can

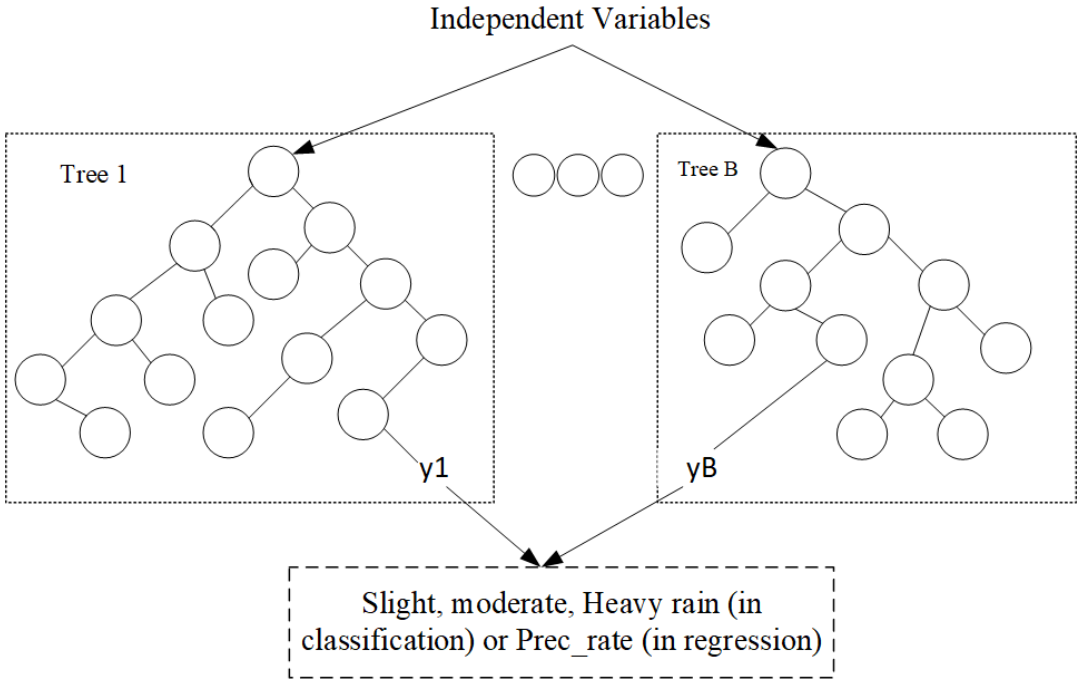


FIGURE 1. Architecture of random forest model.

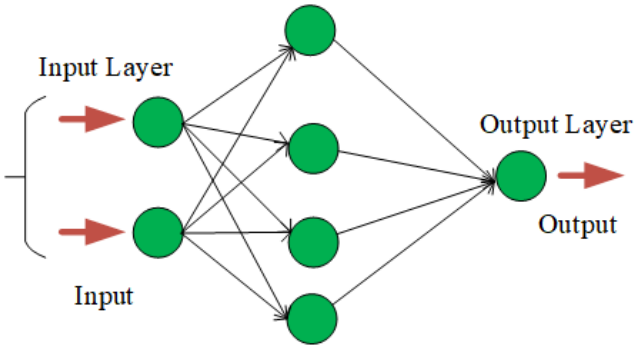


FIGURE 2. A general neural network architecture.

combine several decision trees. Random Forest and Gradient boosting are two examples of these ensemble techniques.

In the case of Random Forest, multiple decision trees are generated by dividing the datasets into random numbers. Subsequently, to avoid overfitting, RF considers each decision tree individually and averages the regression output to provide predictions. Gradient boosting, on the other hand, creates decision trees repeatedly so that the subsequent decision tree corrects the error in the prior tree [1]. In our example, the output is the precipitation rate, and the process is repeated until the final tree is reached.

To reduce the residue, a new tree is iteratively added to the model beginning with the F_O (x), of the model. The gradient boosting concept is depicted schematically in figure 3. The following example shows how to feed input (x) into F_O (x), and pass the output to the next tree.

F. EXTREME GRADIENT BOOSTING

XGBoost, an acronym for extreme gradient boosting, is a particular application of the Gradient Boosting technique that use more precise approximations to determine the optimal tree model [10]. To forecast a target variable y_i, XGBoost is implemented for the supervised machine learning task, which involves data with numerous aspects of x_i. Due to its speed and prediction accuracy, the majority of authors utilize it for various regression and classification tasks.

This method is based on a combination of tree learning algorithm and a linear model. XGBoost is computed concurrently on a single CPU. Due to this fact, it is faster than other gradient descent algorithms. Furthermore, XGBoost is a powerful algorithm that generates consistent predictions due to its efficient memory use and rapid learning speed via parallel and distributed computing [31].

In our analysis, XGBoost was the best regression model for the research problem. Therefore, being the best learning model, it was considered to predict precipitation amounts for weather stations around the Lake Victoria basin. The architecture for XGBoost model is presented in figure 4.

IV. ANALYSIS OF MACHINE LEARNING ALGORITHMS IN THE PREDICTION OF SHORT-TERM RAINFALL AMOUNTS

In this Section, we start by briefly presenting the data description of the variables used in the study. We then present the tools and software used in implementation. Furthermore, the weather data cleaning and correlation analysis is presented to justify our data which was used in the model analysis. We then follow with an analysis of individual

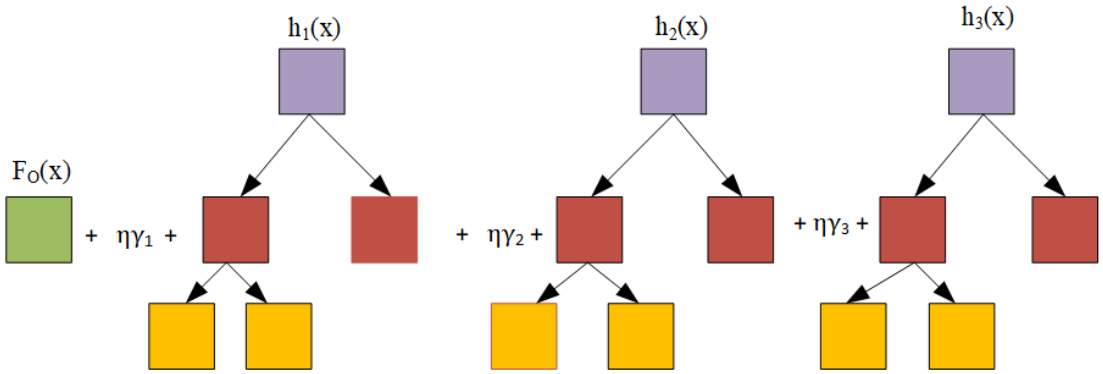


FIGURE 3. The schematic structure of gradient-boosted trees.

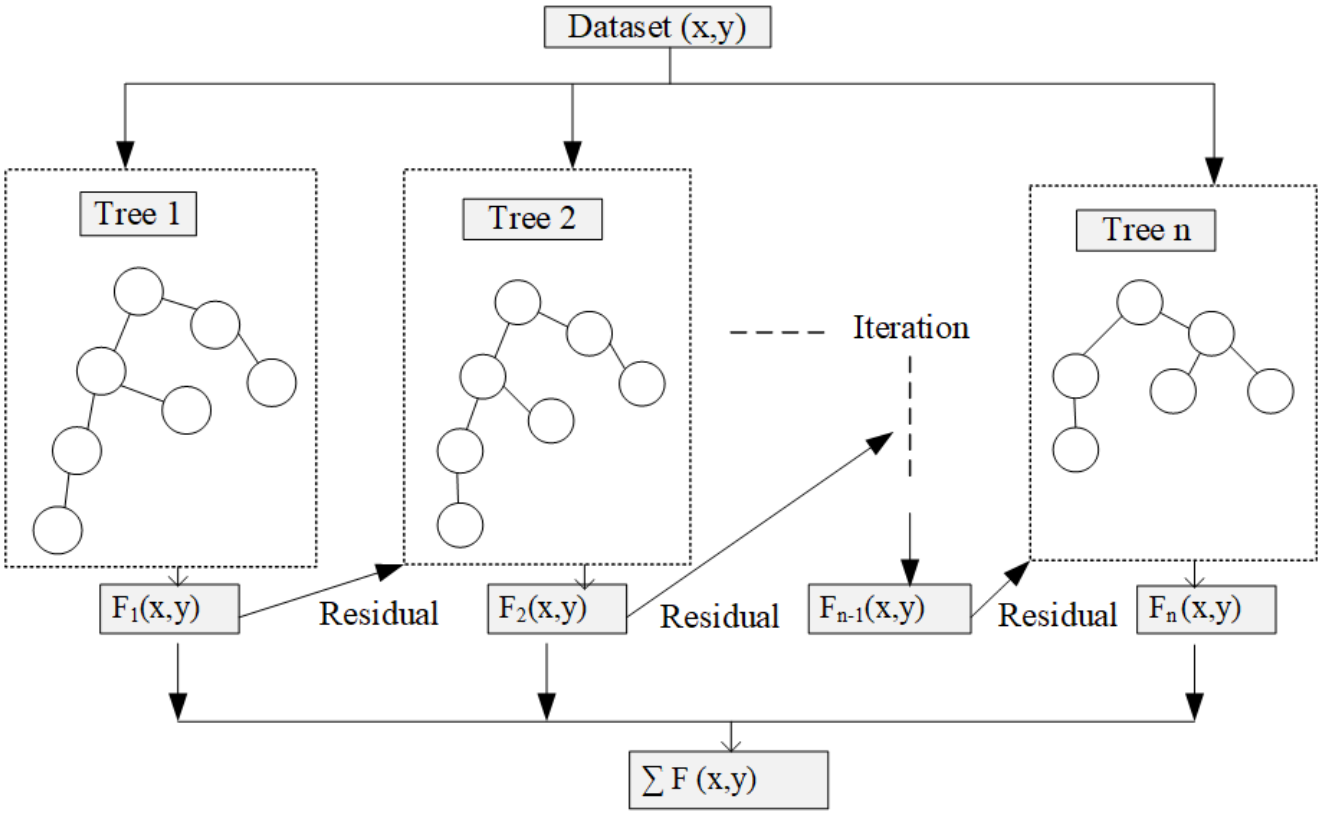


FIGURE 4. The structure of XGBoost model.

ML models used to predict short-term rainfall amounts using our published weather dataset. The entire workflow of the experimentation process used for our ML model analysis is presented in figure 5. The ultimate product of the analysis is to identify the best ML model which we refer to as the most generalizable model for short-term rainfall prediction amounts based on our Uganda's Lake Victoria basin weather dataset.

A. DATA DESCRIPTION

In this study, we used a three (3)-year time period dataset from January 2020 to December 2022. To forecast the amount of precipitation each hour, pre-processing was done on a total of 7 weather stations that cover weather stations around Lake Victoria basin in Uganda, Kenya and Tanzania. The description and the abbreviations for the variables used in this study are shown in table 3. However, the variables Date, P0hPa, and Precmm included in table 3 were not used in the prediction. That said, the date variable will be used in the time series classification task. Also, for the Kampala weather station that was part of the Lake Victoria basin dataset, air pressure at the elevation of the station (P0hPa) was not measured. This variable often has less of an effect on rainfall in low-altitude locations (like Kampala) than it does in high-altitude locales [3].

Furthermore, the data representation of Precmm was recorded in different hourly intervals of 3hrs, 6hrs, 8hrs and more. Therefore, to harmonize the hours uniformly we

converted the different hourly intervals into hourly (per hour). This allowed us to create a new variable precipitation rate per hour which is being investigated in this work; for further information on these conversions, refer to our published article [21], which demonstrates the applied procedures. The basic information about the gauge stations used in this study are shown in table 4. The location of the meteorological weather stations used in this investigation is shown on a map in figure 6.

B. TOOLS AND SOFTWARE USED IN IMPLEMENTATION

The study used Google Colab, a cloud-based Jupyter notebook environment [65]. This platform offers pre-installed libraries, GPU support, and collaborative features to effectively implement and compare NNR, SVR, RF, GBR, LASSO and XGBoost models for precipitation amount prediction around the Lake Victoria basin. Google Colab's cloud-based computing does not have the need for local hardware resources, which makes it a cost-effective and accessible tool for ML tasks. The following Python programming language modules like NumPy, Pandas, Matplotlib, and scikit-learn were used [61]. Numpy objects are primarily used to create arrays or matrices that can be applied to ML models while Pandas provide platforms for data manipulation and analysis purposes. In addition, scikit-learn [65] was utilized for regression tasks among other tasks. Furthermore, deep learning frameworks like PyTorch and TensorFlow provided access to training of neural network models. Matplotlib and Seaborn allowed us to create interactive visuals.

1) DATA CLEANING

During the study, data cleaning processes was done. This is detailed in our published data article [21]. The exploratory data analysis on the dataset was done to maximize the validity of future outcomes, which is critical in machine learning models [18]. The analysis aids in the search for anomalies in the data, the detection of feature correlations, and the search for missing values, all of which help to improve the output of machine learning models.

2) CHARACTERISTICS OF THE UTILIZED WEATHER DATA

In this study, the Augmented Dickey Fuller test (ADF) was employed to evaluate the degree of stationarity of our meteorological dataset for the utilized weather stations. The ADF test was given a significant level of alpha =0.05, and the ADF statistics and critical values (1%, 5% and 10%) were computed. Thus, the P-value for all of the features employed was much less than alpha =0.05, indicating that the data is predictable. Table 5 displays the ADF statistics value for the combined Lake Victoria datasets in Uganda.

3) CORRELATION ANALYSIS

A descriptive analysis of the variables was performed after data pre-processing. The relationship between the variable precipitation rate per hour and the other independent features was specifically investigated. However, because the dependent variable precipitation rate per hour and the independent features in this study did not have a strong correlation, statistical regression methods were not considered for further experiments. Machine learning regression models, on the other hand, were utilized to address the non-linear nature of our weather datasets for the research area. Figure 7 depicts a correlation analysis of the variables considered.

4) DATA NORMALIZATION

The original value of the feature $S(x_i)$ is normalized and the result expressed as;

$$S_n(x_i),\ i = (1, 2 \ldots .., N) \tag{3}$$

$$S_n(x_i) = \frac{(x_i - \bar{x})}{std(x)} \tag{4}$$

where (x_i) is the original value of the features, $\bar{x}$ is the mean (average value of the features across all data points) and std(x) is the standard deviation of the features.

V. ANALYSIS OF REGRESSION MODELS

In this section, we will first of all present the hyperparameters used for optimization of each model and then proceed to present the results.

A. GRADIENT BOOSTING REGRESSION

The Gradient boosting algorithm is a supervised machine learning algorithm that was developed by Friedman [19]. It has proved to be one of the dependable methods for handling complex datasets. The GBR and XGBoost models belong to the same family of gradient descent algorithms. The best hyperparameters for gradient boosting are the number of estimators, the learning rate, and the maximum depth [41]. To determine the ideal values to employ in the boosting model, the grid search algorithm looked through all of the stated values for each parameter and the hyperparameters for GBR are discussed here below;

Number of estimators: Gradient boosting operates based on the principle of decision trees. The n_estimators define the number of sequential trees of the model. Learning rate: Learning rate is an important hyperparameter in gradient boosting. The learning rate scales the contribution of each tree. It's a constant value that measures every new tree that is adjusted. Maximum depth: This is the maximum depth of a tree. Max_depth is used to control overfitting as higher depths allow the model to learn relations that are specific to a particular sample. Minimum sample split: The min_sample_split defines the minimum number of observations that are required in a node required for splitting.

The tuned hyperparameters for the GBR model are given in table 6 along with the best resulting parameter values. These hyperparameters collectively aim for a balanced model that is robust, interpretable, and less prone to overfitting. These specific choices align with considerations for controlling model complexity and generalization [42].

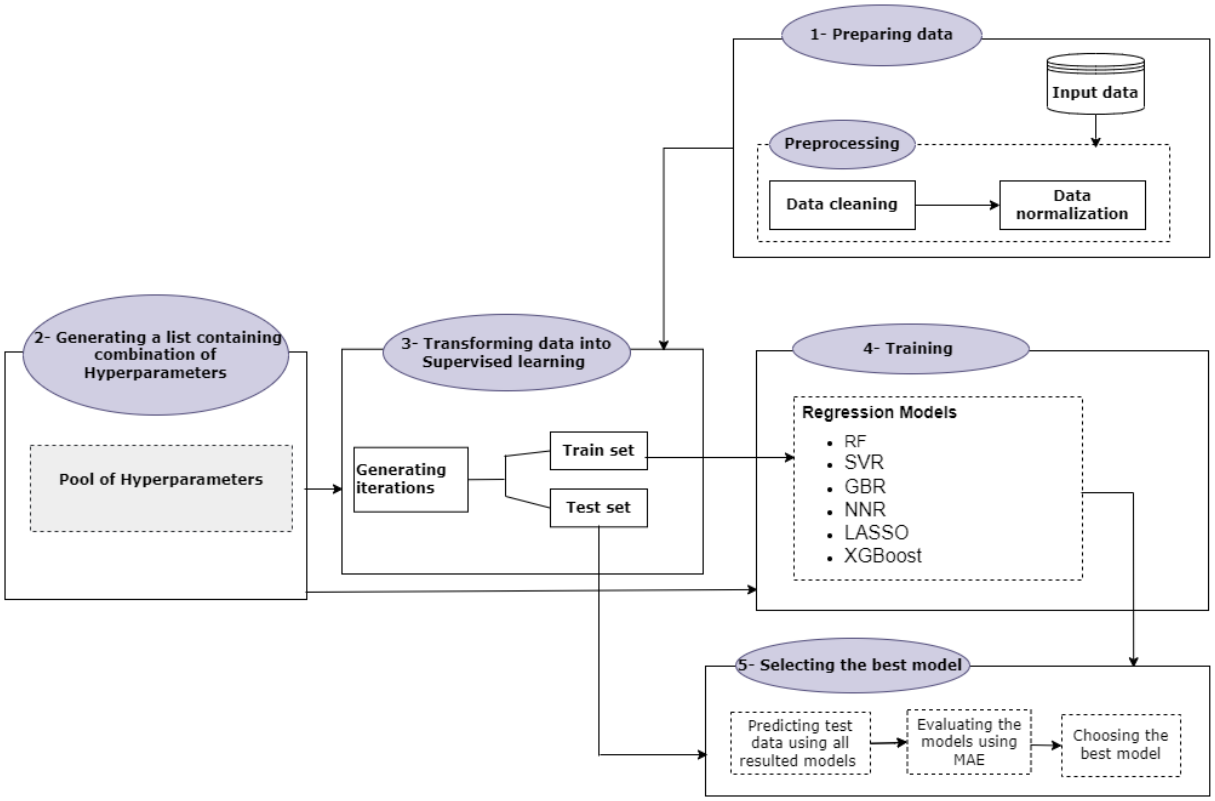


FIGURE 5. Experimental workflow for machine learning model analysis.

TABLE 3. Input variable description.

Feature Abbreviation	Description	Unit of measure	Data period
Date	Day, month, year, and time (hours)	Date	2020-2022
TC	Air temperature at 2 m above ground level	$^{\circ}C$	2020-2022
TDC	Dew point temperature at 2 m above ground level	$^{\circ}C$	2020-2022
Hr	Relative humidity	%	2020-2022
ddd	Wind direction	Direction	2020-2022
ffkmh	Wind speed	km/h	2020-2022
P0hPa	Air pressure at an elevation of the station	hpa	2020-2022
Precmm	Precipitation	mm	2020-2022
Prec_rate	Precipitation rate per hour	mm	2020-2022
Nt	Total cloud cover	oktas	2020-2022
Nh	Cloud cover by high-level cloud fraction	oktas	2020-2022
HKm	Height of the cloud base	km	2020-2022
Viskm	Visibility	km	2020-2022

B. RANDOM FOREST REGRESSION

Random Forest was also one of the models used in this study. It is an ensemble supervised learning model that predicts by selecting individual multiple decision tree regressors at random see section III-A. Random Forest and Gradient boosting regression share some of the hyperparameters. The differences is in the specific tuning of the parameters and their interpretation. The hyperparameters for RF are n_estimators, max_depth, min_samples_split, and min_samples_leaf. The description of these hyperparameters is discussed in section V-A.

The selection of the best hyperparameter values suggests a collective attempt to strike a compromise between generalization and model complexity. The ranges given here

TABLE 4. Information on gauge stations where weather data was obtained.

WMO ID	Station Names	Longitude	Latitude	Altitude
63680	Kampala	32.6166	0.3166	1144
63705	Entebbe Airport	32.4500	0.0500	1155
63682	Jinja	33.1833	0.4500	1175
63708	Kisumu	34.7679	-0.0917	1157
63756	Mwanza	32.9170	-2.4670	1146
63729	Bukoba	31.8120	-1.3321	1137
63733	Musoma	33.8000	-1.5000	1147

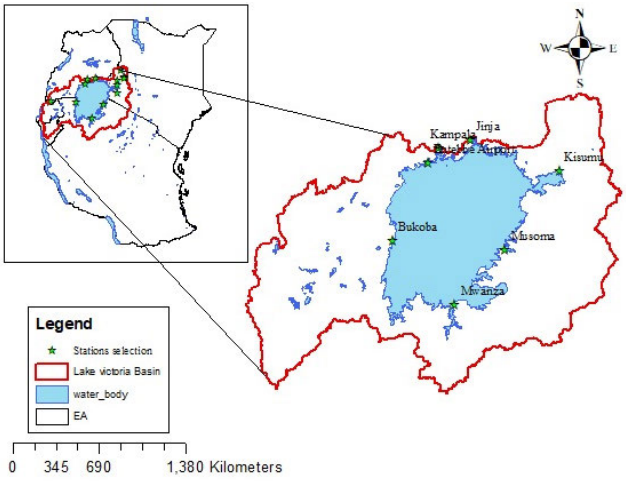


FIGURE 6. Location of weather stations around lake victoria basin.

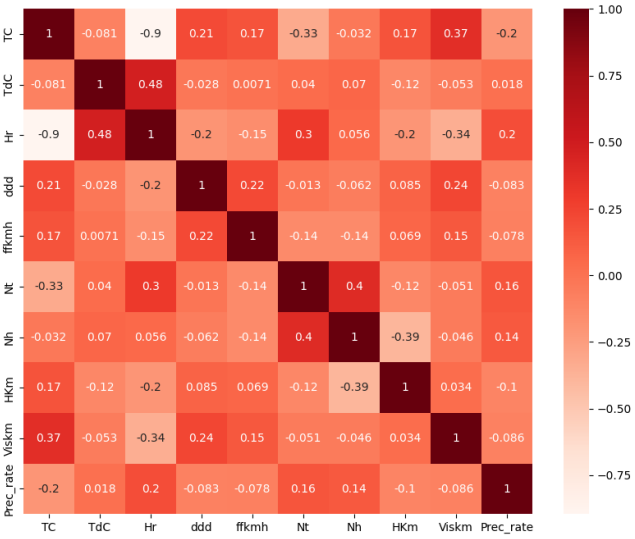


FIGURE 7. Correlation matrix of lake victoria weather data in Uganda.

were sufficient for tuning an RF model [43]. In addition to restrictions on node sizes and tree depth, the moderate number of estimators indicates a desire for a robust model that is less prone to overfitting. The maximum depth of 10 provide the trees more freedom to identify complex patterns in the data. This technique is used to prevent minor, noisy splits in the trees being aligned with the supplied values for min_samples_split and min_samples_leaf. This helps maintain the stability of the model and keeps excessively complicated decision boundaries from forming.

These put into consideration, the selection of these best parameter values results in an RF model that strikes a compromise between resilience and accuracy. The results of the tuned hyperparameters for RF along with best resulting parameter values is shown in table 7.

C. SUPPORT VECTOR REGRESSION

Support Vector Regression is a supervised machine learning approach for classification and regression tasks that operates on the same premise as Support Vector Machines [14]. SVR selects the appropriate hyper-plane in order to maximize margin and reduce error. SVR has several hyperparameters that can be tuned to optimize its performance. Here are some key hyperparameters for SVR; Regularization (C), error sensitivity (ϵ), and kernel (linear, rbf and poly) [40].

Regularization (C): This is the parameter that controls the trade-off between having a smooth decision boundary and fitting the training data [44]. Normally a smaller (C) encourages a smoother boundary, while a larger (C) allows the model to fit the training data more closely. Epsilon (ϵ): Specifies the margin of tolerance where the model is considered to have met the objective. It is particularly relevant in the epsilon-insensitive loss function. Kernel coefficient (gamma): Relevant for the rbf kernel. It determines the shape of the decision boundary. A smaller gamma leads to a more generalized boundary, while a larger gamma can result in a more complex and tighter boundary.

The regularization value (C) of 10 indicates that a higher training error can be tolerated in exchange for a more complex model. In this case where weather data is complex this can be the solution. But thorough validation is required to guarantee that overfitting is kept under control. It is clear that a model that is sensitive to variations within a short-range surrounding the predicted values is desired by using a modest Epsilon ϵ of 0.1. Albeit evaluating the model's performance on unseen data is crucial, this might be appropriate when the emphasis is on precisely capturing the training data.

One of the flexible options for capturing non-linear interactions is the rbf kernel. It helps with handling complex interactions since it enables the SVR model to adjust flexibly to complex patterns in the data. To make sure the SVR model performs well when applied to previously unseen data, it is essential to cross-validate the model's results using these hyperparameters on a different validation set [33]. Table 8 provides a detailed information on the tuned parameters for SVR model.

D. NEURAL NETWORK REGRESSION

Neural Network Regression (NNR) is a type of supervised learning algorithm that uses Artificial Neural Networks to model and predict continuous target variables [45]. For the regression task, the neural network is trained on a

TABLE 5. ADF test results on the time series weather data for selected stations in Uganda.

Variable	ADF statistics	P-value	Critical values		
			1%	5%	10%
TC	-9.985	2.060x10¯17	-3.43471	-2.86347	-2.56779
TdC	-6.717	3.5478x10¯9	-3.43126	-2.86194	-2.56698
Hr	-8.703	3.750x10¯14	-3.43126	-2.86194	-2.56698
ddd	-7.969	2.811x10¯12	-3.43126	-2.86194	-2.56698
ffkmh	-5.880	3.086x10¯7	-3.43126	-2.86194	-2.56698
Nt	-17.044	8.147x10¯30	-3.43126	-2.86194	-2.56698
Nh	-5.6376	1.053x10¯6	-3.43126	-2.86194	-2.56698
HKm	-8.5349	0.0	-3.43126	-2.86194	-2.56698
Viskm	-3.9737	0.00155	-3.43126	-2.86194	-2.56698
Prec_rate	-36.340	-4.457x10¯8	-3.43126	-2.86194	-2.56698

TABLE 6. The hyperparameters for the best resulted GBR model.

Hyperparameter	Parameter values	
	Different values	Best values
Number of estimators	[100, 200, 300, 400, 500]	400
Learning rate	[0.01, 0.05, 0.1, 0.3]	0.01
Maximum depth	[3, 5, 7, 9]	3
min_samples_split	[2, 5, 10]	5
min_samples_leaf	[1, 2, 4]	2

TABLE 7. The hyperparameters for RFR model.

Hyperparameter	Parameter values	
	Different values	Best values
Number of estimators	[100, 200, 300]	100
Maximum depth	[10, 20, 30]	10
min_samples_split	[2, 5, 10]	10
min_samples_leaf	[1, 2, 4]	4

TABLE 8. The hyperparameters for SVR model.

Hyperparameter	Parameter values	
	Different values	Best values
Regularization (C)	[0.1, 1.0, 10.0]	10
Epsilon (ϵ)	[0.1, 0.2, 0.3]	0.1
kernel	[linear, rbf, poly]	rbf

dataset containing input features and corresponding continuous target values. Table 9 represents the hyperparameters values of NNR model. The hyperparameters considered for NN model were batch_size, epochs, learning_rate, and hidden_layer_units [46] as explained here;

Batch Size: This determines the number of samples used in each iteration. Epochs: Decide the number of times the entire dataset is passed forward and backward through the network. Learning Rate: The learning rate is set to control the step size during optimization. Hidden Layer Structure:

TABLE 9. The hyperparameters for NNR model.

Hyperparameter	Parameter values	
	Different values	Best values
Batch_size	[16, 32, 64]	32
Epochs	[50, 100]	100
Learning_rate	[0.001, 0.01, 0.1]	0.01
Hidden_layer_units	[(64, 32), (128, 64), (256, 128)]	(128, 64)

Experiment with the number of hidden layers and the number of neurons in each layer.

The batch size of 32 means that the model updates its weights after processing 32 samples. The number of epochs represents the number of times the entire training dataset is passed forward and backward through the neural network. The 100 epochs is a moderate value that the model has learned the underlying patterns in the data to a satisfactory level [47]. While a learning rate of 0.01 is a reasonable starting point for model development [37]. In addition, a configuration of two hidden layers with 128 units in the first layer and 64 units in the second represents a moderate-sized network that can capture some complexity in the data while potentially avoiding overfitting.

E. EXTREME GRADIENT BOOSTING

Just like the GBR, XGBoost is also an ensemble supervised learning model as discussed in [10]. The tuned XGBoost hyperparameters are the number of estimators, learning rate, maximum depth, reg_alpha, reg_lambda, and optimization technique. The model's performance is improved by tuning the hyperparameter. The tuned parameters for XGBoost [31], [48] are discussed further as follows:

The number of estimators refers to the number of boosting rounds or trees that must be built. A greater number may result in improved performance, but it also increases the risk of overfitting. The learning rate governs how much each

TABLE 10. The hyperparameters for the best resulted XGBoost model.

Hyperparameter	Parameter values	
	Different values	Best values
Number of estimators	[100, 200, 300, 400, 500]	200
Learning rate	[0.01, 0.05, 0.1, 0.3, 0.5]	0.1
Maximum depth	[3, 5, 7, 9, 11]	5
Reg_alpha	[0.1, 0.3, 0.5]	0.1
Reg_lambda	[0.1, 0.3, 0.5]	0.1

TABLE 11. The hyperparameters for LASSO model.

Hyperparameter	Parameter values	
	Different values	Best values
Alpha	[0.001, 0.01, 0.1, 1, 10]	0.1

tree contributes to the final forecast. Lower values make the model more resilient, but they may necessitate more trees. Maximum depth defines each of the tree's maximum depth. Deeper trees can capture more complicated patterns, however, they may over-fit. Reg_lambda is the weights L2 regularization term. It penalizes heavy weights and can assist prevent overfitting. Reg_alpha is similar to reg_lambda, but it penalizes weights with high absolute values.

The obtained value of 200 is reasonable and strikes a balance between model complexity and computational efficiency [49]. Increasing the number of estimators improves the performance of the model, but it comes at the cost of longer training times. The selected learning rate value of 0.1 is a common starting point and it's not too large to cause convergence issues [50]. The L1 and L2 regularization parameters help to avoid overfitting of the model which fit well within our model assumptions [51]. Table 10 demonstrates the hyperparameter values defined for XGBoost model.

F. LEAST ABSOLUTE SHRINKAGE AND SELECTION OPERATOR

LASSO regression in scikit-learn has a few hyperparameters that you can tune to optimize the model's performance. The primary hyperparameter for LASSO is the regularization strength, represented by the parameter alpha [40]. Here, α is the regularization strength which you can experiment with different values of alpha to find the optimal level of regularization for a specific problem. In our example, the best value was 0.1 as shown in table 11.

The value of α determines the strength of the regularization. For the case of our study, $\alpha = 0.1$ implies a moderately strong L1 regularization. It allows for some shrinkage of less important features towards zero, effectively leading to feature selection.

G. SELECTING THE BEST MODEL

The optimised models utilized in the previous phase were evaluated using MAE and RMSE to choose the best performing model. The best optimized model out of all the trained models for predicting precipitation amounts around the Lake Victoria basin in Uganda, Kenya, and Tanzania was XGBoost regression model.

1) EVALUATION METRICS

Generally, when building a machine learning model, the evaluation of its performance is an indispensable step. For this study MAE and RMSE were the adopted evaluation indicators. MAE is a measure of the average absolute differences between predicted and actual values. RMSE is a measure of the average magnitude of the errors between predicted and actual values. Consequently, the lower the MAE and RMSE the better the model's performance. However, the MAE is deemed more appropriate because it is less prone to outliers [52]; and provides a straightforward measure of how well a regression model is performing by summarizing the average absolute errors between predicted and actual values [19]. The MAE and RMSE results are presented in table 12. The MAE and RMSE are defined as follows:

$$MAE = \frac{1}{n}\sum_{x=1}^{n}|y_i - \widehat{y_i}| \tag{5}$$

$$RMSE = \sqrt{\frac{1}{n}\sum\nolimits_{i=1}^{n}(y_i - \widehat{y_i})^2} \tag{6}$$

where n is the total number of data points, y_x is the actual (observed) value of the target variable for the i^{th} data point and $\widehat{y_i}$ represent the predicted value of the target.

2) SPECIFICATIONS OF XGBoost

XGBoost approximates the loss function with Taylor expansions, giving the model superior trade-off bias and variance while requiring fewer decision trees to achieve higher accuracy [11]. Suppose the given sample set has **n** samples and **m** features, it can be represented as;

$$D = \{(x_i, y_i)\}(|D| = n, x_i \epsilon R^{\mathrm{m}}, y \epsilon R) \tag{7}$$

$$\widehat{y_i} = \sum_{k=1}^{K} f_k(x_i), f_k \epsilon F \tag{8}$$

F is the set of decision trees:

$$F = \{f(x) = W_{q(x)}\}(q : R^{\mathrm{m}} \rightarrow T, w, \epsilon R^{\mathrm{T}}) \tag{9}$$

where $f(x)$ is one of the trees, $W_{q(x)}$ is the weight of the leaf nodes, T is the number of leaf nodes, and q represents the structure of each tree, which maps the samples to the corresponding leaf nodes.

Here is the feature value and is the actual value. The algorithm sums the results of the tree to the final predicted value, $\hat{y}_i$ denoted as;

$$\widehat{y_i} = \sum_{k=1}^{K} f_k(x_i), f_k \epsilon F \tag{10}$$

F is the set of decision trees:

$$F = \{f(x) = W_{q(x)}\}(q : R^{\mathrm{m}} \rightarrow T, w, \epsilon R^{\mathrm{T}}), \tag{11}$$

TABLE 12. Comparison of MAE and RMSE with benchmark techniques using Uganda weather data.

Method	Test		Ranking	
	MAE	RMSE	MAE	RMSE
RFR	0.12183	0.42925	5	5
SVR	0.05888	0.12302	3	3
NNR	0.01018	0.05952	2	2
LASSO	0.06610	0.26626	4	4
GBR	0.12624	0.43931	6	6
XGBoost	**0.00616**	**0.04439**	1	1

where $f(x)$ is one of the trees, $W_{q(x)}$ is the weight of the leaf nodes, T is the number of leaf nodes, and q represents the structure of each tree, which maps the samples to the corresponding leaf nodes.

The objective function of XGBoost model can be divided into error function term $\mathcal{L}$ and model complexity function term, Ω [44]. The objective function can be written as:

$$Obj = \mathcal{L} + \Omega \tag{12}$$

Therefore, the predicted value of XGBoost is the sum of the values of the leaf nodes of each tree. The goal of this model is to learn this K tree, so minimizing the following objective function.

$$Obj = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \gamma^T + \frac{1}{2}\lambda \sum_{j=1}^{T} W_j^2 \tag{13}$$

VI. COMPARISON OF MODEL PERFORMANCES

In this section we compare the performance of the selected models to determine the best model.

A. EVALUATING XGBoost

Literature-backed evidence suggests that there is no dominant forecasting model for different machine learning applications [53]. We also acknowledge that several researchers have used a variety of forecasting methodologies using different performance metrics as discussed in table 2 above. Thus, the predictive power of XGBoost model is compared to that of other cutting-edge models, including Random Forest regressors, Support Vector regression, Neural Network regression, Gradient Boost regression, and Lasso regression. It is vital to highlight that for each benchmark technique, we looked for ideal parameters that would allow the strategy to perform the best.

Table 12 and figures 8 and 9 compares the performance of the proposed method to the benchmark techniques on Uganda's weather data for the Lake Victoria basin. Notably, the results illustrate how the corresponding approach performs in terms of MAE and RMSE on test set performance.

B. COMPARISON OF XGBoost WITH BENCHMARK TECHNIQUES IN KENYA AND TANZANIA

The developed XGBoost model for Uganda was applied to datasets for stations located around the Lake Victoria basin in Kenya and Tanzania. The aim was to test the model's generalizability in regions with comparable weather patterns [54], [58], [60]. Consequently, 20% of data was drawn from Kisumu Kenya weather station and another 20% from Tanzania weather stations. The experiments were done independently. From the findings, XGBoost performed the best in both nations for predicting precipitation amounts around the Lake Victoria basin. The comparison of the models is shown in table 13 and figures 10 and 11 respectively.

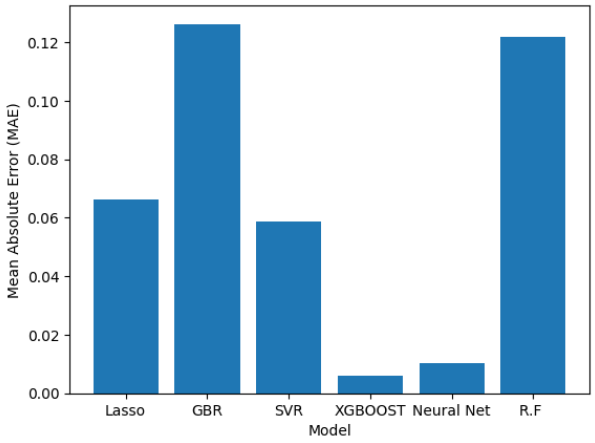


FIGURE 8. Comparison of regression algorithm performance using Uganda weather data.

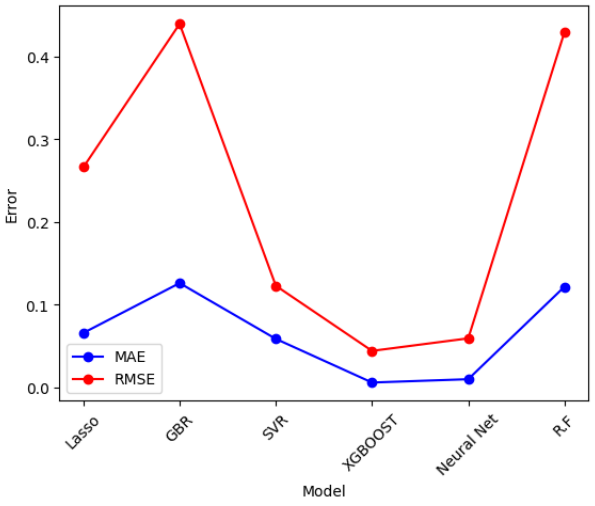


FIGURE 9. MAE across regression algorithms performance using Uganda weather data.

C. XGBoost METHOD VERSUS BENCHMARK TECHNIQUES

In this section, we compare the performance of XGBoost model with the applied machine learning techniques.

1) XGBoost VERSUS RANDOM FOREST REGRESSOR

Random Forest is one of the best models utilized for various applications including weather prediction [34]. However, its performance was not appropriate for our data. Considering

TABLE 13. Comparison of MAE and RMSE with benchmark techniques using Kenya and Tanzania weather data.

		Performance Metrics	
Stations	Model	MAE	RMSE
Kenya	RFR	0.478242	0.702832
	SVR	0.143649	0.313051
	NNR	0.426446	0.623602
	LASSO	0.424491	0.592471
	GBR	0.315397	0.472994
	XGBoost	**0.018452**	**0.227948**
Tanzania	RFR	0.233441	0.842882
	SVR	0.081293	0.487932
	NNR	0.144132	0.702039
	LASSO	0.233208	0.759692
	GBR	0.055811	0.462857
	XGBoost	**0.005862**	**0.213369**

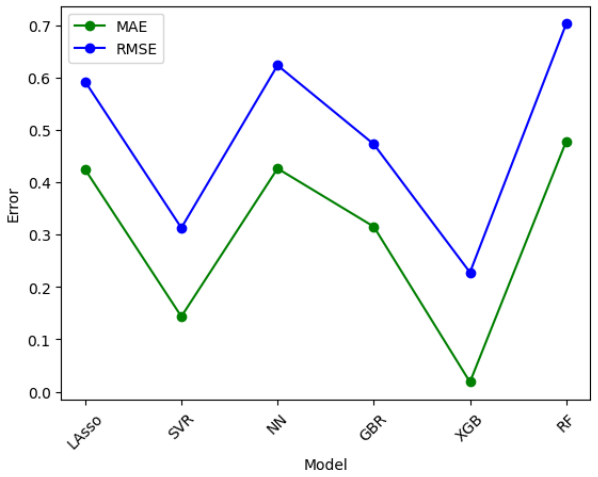


FIGURE 10. Comparison of regression algorithm performance using Tanzania weather data.

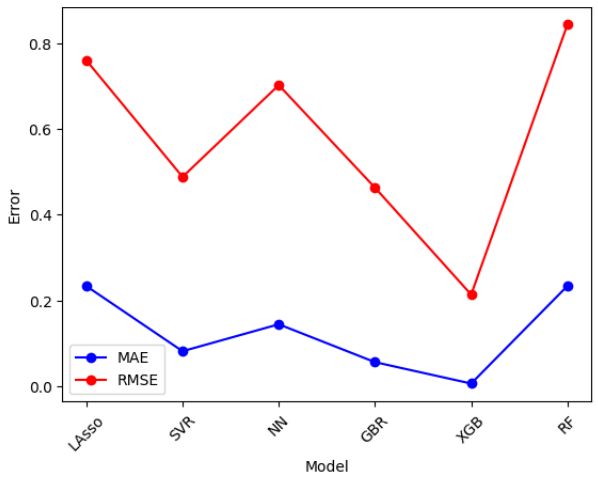


FIGURE 11. Comparison of regression algorithm performance using Kenya weather data.

the MAE of 0.12183, it did not perform well compared to its competitors.

2) XGBoost VERSUS SUPPORT VECTOR REGRESSION

Support vector regression has showed good prediction performance in a variety of disciplines, including weather prediction [55] and solar power forecasting [39], among others. Kernel based SVR model is less prone to overfitting [56]. These characteristics make this approach more feasible than Artificial Neural Networks. For our dataset, SVR came close to predicting precipitation amounts over the Lake Victoria in Uganda. In comparison to other models, it was the third-best model with an MAE of 0.058885.

3) XGBoost VERSUS NEURAL NETWORK REGRESSION

To solve non-linear problems in time series weather data, neural networks (NN) have been widely used [37]. Despite the fact that ANNs face some obstacles in weather modeling, different ANNs can handle the problem of local minima that plagues neural network topologies. When compared to XGBoost model, NN performed second best with MAE of 0.01018. The effectiveness of NN is not surprising, as it often performs well with non-linear data [36], [37], such as the meteorological data utilized in this investigation.

4) XGBoost VERSUS LASSO

Least Absolute Shrinkage and Selection Operator is one of the commonly used regression techniques. It can also improve the precision and lessen the variability of techniques to linear regression [40]. When used on our dataset, LASSO performed poorly compared to its competitors. In terms of MAE on the test data, LASSO did not perform well compared to its rivals SVR, XGBoost, and NNR in our analysis.

5) XGBoost VERSUS GRADIENT BOOST REGRESSION

Friedman's gradient boosting technique [19] is a supervised learning method. It has proven to be a very reliable method for many complex datasets. GBR employs iteration of a collection of weak learners to generate one strong learner using an appropriate loss function [10]. However, when applied to our dataset, GBR was the worst performing model when compared to its competitors. GBR models only perform well when applied to smaller datasets [44]. Albeit, GBR and XGBoost belong in the same family of gradient boosting. XGBoost [10] is stronger than GBR because of several factors including speed, regularization abilities, and handling missing values.

D. STATISTICAL COMPARISONS OF THE APPLIED MODELS

The statistical tests were performed to establish whether the proposed method's performance is noticeably superior to that of the competing approaches. As a result, we ran experiments to demonstrate statistical significant differences between the outputs of the RFR, SVR, NNR, LASSO, GBR and XGBoost techniques while performing k-fold cross-validation on the predictive models. The prediction algorithm has a size k of 5. The findings of this experiment are expressed in terms of MAE. Table 14 shows that XGBoost achieved lower

TABLE 14. Comparison of RFR, SVR, NNR, LASSO, GBR, and XGB with k-fold cross validation (k=5).

K-fold values	RFR	SVR	NNR	LASSO	GBR	XGBoost
1	0.099690	0.064194	0.0196971	0.073976	0.124082	0.003839
2	0.106142	0.058574	0.0070363	0.080993	0.111952	0.006504
3	0.088517	0.085452	0.0145868	0.099539	0.145670	0.024326
4	0.154504	0.072535	0.0117364	0.094721	0.125953	0.017161
5	0.202481	0.060914	0.0250217	0.074744	0.105899	0.005515
Total out of sample	0.130267	0.068334	0.0156157	0.084795	0.122711	0.006167

TABLE 15. Wilcoxon signed rank test difference between GBR with XGBoost.

K-fold values	GBR	XGBoost	Difference	Positive	|Difference|	Rank	Signed Rank
1	0.12408	0.00383	-0.12024	-1	0.12024	4	-4
2	0.11195	0.00650	-0.10545	-1	0.10545	2	-2
3	0.14567	0.02432	-0.12134	-1	0.12134	5	-5
4	0.12595	0.01716	-0.10879	-1	0.10879	3	-3
5	0.10589	0.00551	-0.10038	-1	0.10038	1	-1

Positive sum: 0; Negative sum: 15; Test Statistics: 0. The number of negative differences is greater than the positive differences, which indicates XGBoost is working better than GBR.

TABLE 16. Wilcoxon signed rank test difference between LASSO with XGBoost.

K-fold values	LASSO	XGBoost	Difference	Positive	|Difference|	Rank	Signed Rank
1	0.07397	0.00383	-0.07014	-1	0.07014	2	-2
2	0.08099	0.00650	-0.07449	-1	0.07449	3	-3
3	0.09953	0.02432	-0.07521	-1	0.07521	4	-4
4	0.09472	0.01716	-0.07756	-1	0.07756	5	-5
5	0.07474	0.00551	-0.06923	-1	0.06923	1	-1

Positive sum: 0; Negative sum: 15; Test Statistics: 0. The number of negative differences is greater than the positive differences, which indicates XGBoost is working better than LASSO.

MAE than RFR, SVR, NNR, LASSO and GBR prediction algorithms. We adopted the Wilcoxon Signed Rank test applied in work [17] to compare XGBoost against the five machine learning regression techniques.

The employed theory H_0: The machine learning approach outperforms XGBoost if the number of positive differences exceeds the number of negative differences. H_1: If there are more negative differences than positive differences, the XGBoost algorithm outperforms the competing machine learning regression techniques. Tables 15-19 exhibit the Wilcoxon Signed Rank test results of RFR, SVR, NNR, LASSO and GBR with XGBoost.

E. COMPARATIVE ANALYSIS OF XGBoost WITH EXISTING METHODS

The results shown in this section are consistent with state-of-the-art techniques. In this section, we compare the proposed model's results to those of previously utilized rainfall forecast techniques. The findings of our investigations are not different from previous researchers. For example, we have found that weather data is in fact non-linear. In our dataset, there was no correlation found between the dependent variable precipitation rate and independent factors. As such, the present analysis confirms to the findings of studies [15], [18], [20], [37] that say weather data is in fact non-linear. This is a reason non-linear models were used in precipitation amount prediction around the Lake Victoria basin.

Also, we evaluate how well the suggested XGBoost model performs in short-term weather forecasting compared to other state-of-the-art methods. Due to its speed and accuracy of the predictions, XGBoost has done better than other benchmark models in numerous studies. Our XGBoost results are comparable to rainfall forecast models that have been used in the past when we compare the outcomes of the proposed model with theirs. In this study, MAE of 0.00616 for Uganda, 0.01845 for Kenya, and 0.00586 for Tanzania were obtained using the proposed XGBoost model for the Lake Victoria basin.

In contrast to other researchers who have forecasted short-term precipitation using the XGBoost model. Liyew and Melese [10], for example, got an MAE of 3.58 in Ethiopian short-term rainfall prediction. Researchers [37] used the XGBoost model to predict hourly rainfall in the UK and obtained MAE results of 0.05, 0.08, 0.04, 0.07 and 0.07 for the cities of Bath, Bristol, Cardiff, Newport, and Swindon, respectively. XGBoost was one of the regression

TABLE 17. Wilcoxon signed rank test difference between NNR with XGBoost.

K-fold values	NNR	XGBoost	Difference	Positive	|Difference|	Rank	Signed Rank
1	0.019697	0.00383	-0.01586	-1	0.01586	4	-4
2	0.007036	0.00650	-0.00053	-1	0.00053	1	-1
3	0.014587	0.02432	0.00974	1	0.00974	3	3
4	0.011736	0.01716	0.00542	1	0.00542	2	2
5	0.025022	0.00551	-0.01951	-1	0.01951	5	-5

Positive sum: 5; Negative sum: 10; Test Statistics: 5. The number of negative differences is greater than the positive differences, which indicates XGBoost is working better than NNR.

TABLE 18. Wilcoxon signed rank test difference between SVR with XGBoost.

K-fold values	SVR	XGBoost	Difference	Positive	|Difference|	Rank	Signed Rank
1	0.064195	0.00383	-0.06036	-1	0.06036	4	-4
2	0.058574	0.00650	-0.05207	-1	0.05207	1	-1
3	0.085452	0.02432	-0.06113	-1	0.06113	5	-5
4	0.072535	0.01716	-0.05537	-1	0.05537	2	-2
5	0.060914	0.00551	-0.05540	-1	0.05540	3	-3

Positive sum: 0; Negative sum: 15; Test Statistics: 0. The number of negative differences is greater than the positive differences, which indicates XGBoost is working better than SVR.

TABLE 19. Wilcoxon signed rank test difference between RFR with XGBoost.

K-fold values	RFR	XGBoost	Difference	Positive	|Difference|	Rank	Signed Rank
1	0.09969	0.00383	-0.09585	-1	0.09585	2	-2
2	0.106142	0.00650	-0.09964	-1	0.09964	3	-3
3	0.088517	0.02432	-0.06419	-1	0.06419	1	-1
4	0.154504	0.01716	-0.13734	-1	0.13734	4	-4
5	0.202481	0.00551	-0.19697	-1	0.19697	5	-5

Positive sum: 0; Negative sum: 15; Test Statistics: 0. The number of negative differences is greater than the positive differences, which indicates XGBoost is working better than RFR.

models employed by Kanani et al. [18] to forecast the amount of precipitation. The MAE of 0.121 was obtained for the XGBoost model.

Also, XGBoost was used by Anwar et al. [59] to forecast daily rainfall estimations. The RMSE of 2.7 mm and an MAE of 8.8 mm were obtained in their study. Rainfall prediction in Ghana's several ecological zones was shown to be better predicted using XGBoost, RF, and MLP, according to Appiah-Badu et al. [60]. The study by Lawal et al. [65] found XGBoost as the best model with an MAE of 0.042529 and RMSE of 0.05654 in predicting daily rainfall in the Nyando region of Kenya. In addition, Ganapathy et al. [66] identified XGBoost as the best prediction model for daily rainfall prediction in the localized Vellore region of Tamil Nadu, India, for years 2021 and 2022.

However, it is important to highlight that our study applied the Lake Victoria basin weather dataset. As previously noted different authors have used different meteorological data from their countries. The conclusions that we learn from each of these studies is that weather data in locations with comparable weather patterns can be predicted using ML models. In our study, the obtained XGBoost results are comparable to previous works. We can confirm that the applied XGBoost model had acceptable results for predicting precipitation amounts around the Lake Victoria basin in Uganda, Kenya and Tanzania. However, pre-processing, model building, and the kind of datasets all affect how well the models perform.

VII. CONCLUSION AND FUTURE WORK

This study examined various machine learning algorithms for predicting hourly precipitation amounts around the Lake Victoria basin. Six machine learning regression algorithms were trained, tested, and validated using our published weather dataset.

The study revealed that precipitation rate is crucial in determining the amount of rain that falls on an hourly basis. This would be significant for a variety of industries, including agriculture, tourism, aviation, education and engineering. This can lessen the problems associated with high rainfall intensity, such as flood mortality and disease caused by rain.

The relevant environmental features for precipitation amount prediction were experimented using the Pearson correlation coefficient to determine their relationship with the target variable. However, the results showed that the variables hard a weak correlation with the target variable. Therefore,

non-linear machine learning models were used to input the variables used in the study. A comparison of results among the six algorithms; RFR, SVR, LASSO, GBR, NNR and XGBoost was made and the results showed that the XGBoost was a better-suited machine learning regression algorithm for precipitation amount prediction using selected environmental features.

In future works, we intend to do cross station data modeling [58]. This is critical in areas with limited meteorological data where data from areas with similar geographical characteristics can be utilized. Albeit, this can be done with data overtime, the scarcity of meteorological data remains a challenge. We therefore propose to do weather pattern mapping, followed by transfer learning models. However, the current used weather dataset can further be improved to include geographical information such as latitude, longitude and altitude [62]. This should further improve the performance of transfer learning models.

ETHICS STATEMENT

The study does not involve experiments on humans or animals.

CREDIT AUTHORSHIP CONTRIBUTION STATEMENT

Tumusiime Andrew Gahwera: Conceptualization, data curation, methodology, writing–original draft. Odongo Steven Eyobu: Supervision, data curation, writing–review and editing. Mugume Isaac: Supervision, investigation, writing–review and editing.

DECLARATION OF COMPETING INTEREST

The authors declare that there were no known competing commercial interests or personal relationships that could have appeared to influence this work.

ACKNOWLEDGMENT

The author Tumusiime Andrew Gahwera thanks his supervisors: Dr. Odongo Steven Eyobu and Dr. Mugume Isaac for their guidance.

REFERENCES

[1] V. S. Monego, J. A. Anochi, and H. F. de Campos Velho, "South America seasonal precipitation prediction by gradient-boosting machine-learning approach," *Atmosphere*, vol. 13, no. 2, p. 243, Jan. 2022, doi: 10.3390/atmos13020243.

[2] R. Vijayan, V. Mareeswari, P. Mohankumar, and G. G. K. Srikar, "Estimating rainfall prediction using machine learning techniques on a dataset," *Int. J. Sci. Technol. Res.*, vol. 9, no. 6, pp. 440–445, 2020.

[3] R. Opio, G. Sabiiti, A. Nimusiima, I. Mugume, and J. Sansa-Otim, "WRF simulations of extreme rainfall over Uganda's lake Victoria basin: Sensitivity to parameterization, model resolution and domain size," *J. Geosci. Environ. Protection*, vol. 8, no. 4, pp. 18–31, 2020, doi: 10.4236/gep.2020.84002.

[4] D. Ntwali, B. A. Ogwang, and V. Ongoma, "The impacts of topography on spatial and temporal rainfall distribution over Rwanda based on WRF model," *Atmos. Climate Sci.*, vol. 6, no. 2, pp. 145–157, 2016, doi: 10.4236/acs.2016.62013.

[5] J. A. Anochi, V. A. de Almeida, and H. F. de Campos Velho, "Machine learning for climate precipitation prediction modeling over south America," *Remote Sens.*, vol. 13, no. 13, p. 2468, Jun. 2021.

[6] D. Ying, P. Hua, and M. Hao, "Research and application of SMOTE-based method with XGBoost regression prediction," in *Proc. IEEE Int. Conf. Image Process. Comput. Appl. (ICIPCA)*, Aug. 2023, pp. 1737–1740, doi: 10.1109/icipca59209.2023.10257809.

[7] C. Huntingford, E. S. Jeffers, M. B. Bonsall, H. M. Christensen, T. Lees, and H. Yang, "Machine learning and artificial intelligence to aid climate change research and preparedness," *Environ. Res. Lett.*, vol. 14, no. 12, Dec. 2019, Art. no. 124007, doi: 10.1088/1748-9326/ab4e55.

[8] A. Gupta, H. K. Mall, and S. Janarthanan., "Rainfall prediction using machine learning," in *Proc. 1st Int. Conf. Artif. Intell. Trends Pattern Recognit. (ICAITPR)*, Mar. 2022, pp. 1–5, doi: 10.1109/ICAITPR51569.2022.9844203.

[9] A. Samad, V. Gautam, P. Jain, and K. Sarkar, "An approach for rainfall prediction using long short term memory neural network," in *Proc. IEEE 5th Int. Conf. Comput. Commun. Autom. (ICCCA)*, Oct. 2020, pp. 190–195, doi: 10.1109/ICCCA49541.2020.9250809.

[10] C. M. Liyew and H. A. Melese, "Machine learning techniques to predict daily rainfall amount," *J. Big Data*, vol. 8, no. 1, pp. 1–11, Dec. 2021.

[11] H. Chen and L. Chen, "An application of XGBoost algorithm for online transaction fraud detection based on improved sailfish optimizer," in *Proc. 4th Int. Conf. Mach. Learn., Big Data Bus. Intell. (MLBDBI)*, Oct. 2022, pp. 294–299, doi: 10.1109/MLBDBI58171.2022.00064.

[12] K. U. Jaseena and B. C. Kovoor, "Deterministic weather forecasting models based on intelligent predictors: A survey," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 34, no. 6, pp. 3393–3412, Jun. 2022, doi: 10.1016/j.jksuci.2020.09.009.

[13] M. S. Balamurugan and R. Manojkumar, "Study of short term rain forecasting using machine learning based approach," *Wireless Netw.*, vol. 27, no. 8, pp. 5429–5434, Nov. 2021, doi: 10.1007/s11276-019-02168-3.

[14] V. P. Tharun, R. Prakash, and S. R. Devi, "Prediction of rainfall using data mining techniques," in *Proc. 2nd Int. Conf. Inventive Commun. Comput. Technol. (ICICCT)*, Apr. 2018, pp. 1507–1512, doi: 10.1109/ICICCT.2018.8473177.

[15] K. Dutta and P. Gouthaman, "Rainfall prediction using machine learning and neural network," *Int. J. Recent Technol. Eng.*, vol. 9, no. 1, pp. 1954–1961, May 2020, doi: 10.35940/ijrte.a2747.059120.

[16] S. E. Haupt, J. Cowie, S. Linden, T. McCandless, B. Kosovic, and S. Alessandrini, "Machine learning for applied weather prediction," in *Proc. IEEE 14th Int. Conf. e-Sci.*, Oct. 2018, pp. 276–277, doi: 10.1109/ESCIENCE.2018.00047.

[17] D. Endalie, G. Haile, and W. Taye, "Deep learning model for daily rainfall prediction: Case study of Jimma, Ethiopia," *Water Supply*, vol. 22, no. 3, pp. 3448–3461, Mar. 2022, doi: 10.2166/ws.2021.391.

[18] S. Kanani, S. Patel, R. K. Gupta, A. Jain, and J. C.-W. Lin, "An AI-enabled ensemble method for rainfall forecasting using long-short term memory," *Math. Biosci. Eng.*, vol. 20, no. 5, pp. 8975–9002, 2023, doi: 10.3934/mbe.2023394.

[19] A. S. Ponraj and T. Vigneswaran, "Daily evapotranspiration prediction using gradient boost regression model for irrigation planning," *J. Supercomput.*, vol. 76, no. 8, pp. 5732–5744, Aug. 2020, doi: 10.1007/s11227-019-02965-9.

[20] M. Mohammed, R. Kolapalli, N. Golla, and S. S. Maturi, "Prediction of rainfall using machine learning techniques," *Int. J. Sci. Technol. Res.*, vol. 9, no. 1, pp. 3236–3240, 2020.

[21] A. G. Tumusiime, O. S. Eyobu, I. Mugume, and T. J. Oyana, "A weather features dataset for prediction of short-term rainfall quantities in Uganda," *Data Brief*, vol. 50, Oct. 2023, Art. no. 109613, doi: 10.1016/j.dib.2023.109613.

[22] O. M. Adisa, M. Masinde, J. O. Botai, and C. M. Botai, "Bibliometric analysis of methods and tools for drought monitoring and prediction in Africa," *Sustainability*, vol. 12, no. 16, p. 6516, Aug. 2020, doi: 10.3390/su12166516.

[23] C. Castillo-Botón, D. Casillas-Pérez, C. Casanova-Mateo, S. Ghimire, E. Cerro-Prada, P. A. Gutierrez, R. C. Deo, and S. Salcedo-Sanz, "Machine learning regression and classification methods for fog events prediction," *Atmos. Res.*, vol. 272, Jul. 2022, Art. no. 106157, doi: 10.1016/j.atmosres.2022.106157.

[24] U. Vivek Krishna, S. Sanju, S. Sudhakaran, R. Kaladevi, and H. Shanmugasundaram, "Weather prediction using linear regression model," in *Proc. Int. Conf. Comput. Commun. Informat. (ICCCI)*, Jan. 2023, pp. 1–4, doi: 10.1109/ICCCI56745.2023.10128252.

[25] N. Karna and P. C. Roy, "Temperature prediction using regression model," *Adv. Eng. ICT Converg. Proc.*, vol. 10, no. 4, pp. 161–170, Jul. 2021.

[26] M. Navid, "Multiple linear regressions for predicting rainfall for Bangladesh," *Communications*, vol. 6, no. 1, p. 1, 2018, doi: 10.11648/j.com.20180601.11.

[27] B. M. Preethi, R. Gowtham, S. Aishvarya, S. Karthick, and D. G. Sabareesh, "Rainfall prediction using machine learning and deep learning algorithms," *Int. J. Recent Technol. Eng.*, vol. 10, no. 4, pp. 251–254, Nov. 2021, doi: 10.35940/ijrte.d6611.1110421.

[28] A. Garg and H. Pandey, "Rainfall prediction using machine learning," *Int. J. Innov. Sci. Res. Technol.*, vol. 4, no. 5, pp. 56–58, May 2019.

[29] H. Zheng and Y. Wu, "A XGBoost model with weather similarity analysis and feature engineering for short-term wind power forecasting," *Appl. Sci.*, vol. 9, no. 15, p. 3019, Jul. 2019.

[30] X. Deng, A. Ye, J. Zhong, D. Xu, W. Yang, Z. Song, Z. Zhang, J. Guo, T. Wang, Y. Tian, H. Pan, Z. Zhang, H. Wang, C. Wu, J. Shao, and X. Chen, "Bagging–XGBoost algorithm based extreme weather identification and short-term load forecasting model," *Energy Rep.*, vol. 8, pp. 8661–8674, Nov. 2022, doi: 10.1016/j.egyr.2022.06.072.

[31] X. Ma, C. Fang, and J. Ji, "Prediction of outdoor air temperature and humidity using Xgboost," *IOP Conf. Ser., Earth Environ. Sci.*, vol. 427, no. 1, pp. 1–9, 2020, doi: 10.1088/1755-1315/427/1/012013.

[32] M. P. Clark, R. M. Vogel, J. R. Lamontagne, N. Mizukami, W. J. M. Knoben, G. Tang, S. Gharari, J. E. Freer, P. H. Whitfield, K. R. Shook, and S. M. Papalexiou, "The abuse of popular performance metrics in hydrologic modeling," *Water Resour. Res.*, vol. 57, no. 9, pp. 1–15, Sep. 2021, doi: 10.1029/2020wr029001.

[33] B. O. Parlak and H. A. Yavaşoğlu, "Comparison of regression algorithms to predict average air temperature," *Uluslararası Muhendislik Arastirma Gelistirme Dergisi*, vol. 15, no. 1, pp. 312–322, Jan. 2023, doi: 10.29137/umagd.1232020.

[34] T. Zhu, "Analysis on the applicability of the random forest," *J. Phys., Conf. Ser.*, vol. 1607, no. 1, Aug. 2020, Art. no. 012123, doi: 10.1088/1742-6596/1607/1/012123.

[35] P. Chitra and S. Abirami, "A deep learning ensemble model for short-term rainfall prediction," in *Proc. Int. Conf. Wireless Commun. Signal Process. Netw. (WiSPNET)*, Mar. 2022, pp. 135–138, doi: 10.1109/WiSPNET54241.2022.9767163.

[36] E. Hernández, V. Sanchez-Anguix, V. Julian, J. Palanca, and N. Duque, "Rainfall prediction: A deep learning approach," in *Hybrid Artificial Intelligent Systems* (Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Spain: Springer, vol. 9648, 2016, pp. 151–162, doi: 10.1007/978-3-319-32034-2.

[37] A. Y. Barrera-Animas, L. O. Oyedele, M. Bilal, T. D. Akinosho, J. M. D. Delgado, and L. A. Akanbi, "Rainfall prediction: A comparative analysis of modern machine learning algorithms for time-series forecasting," *Mach. Learn. Appl.*, vol. 7, Mar. 2022, Art. no. 100204, doi: 10.1016/j.mlwa.2021.100204.

[38] N. Khan, D. A. Sachindra, S. Shahid, K. Ahmed, M. S. Shiru, and N. Nawaz, "Prediction of droughts over Pakistan using machine learning algorithms," *Adv. Water Resour.*, vol. 139, May 2020, Art. no. 103562, doi: 10.1016/j.advwatres.2020.103562.

[39] M. Abuella and B. Chowdhury, "Solar power forecasting using support vector regression," in *Proc. Int. Annu. Conf. Amer. Soc. Eng. Manag.*, Mar. 2017, pp. 1-6.

[40] G. V. Sajan and P. Kumar, "Forecasting and analysis of train delays and impact of weather data using machine learning," in *Proc. 12th Int. Conf. Comput. Commun. Netw. Technol. (ICCCNT)*, Jul. 2021, pp. 1–8, doi: 10.1109/ICCCNT51525.2021.9580176.

[41] D. Sangani, K. Erickson, and M. A. Hasan, "Predicting zillow estimation error using linear regression and gradient boosting," in *Proc. IEEE 14th Int. Conf. Mobile Ad Hoc Sensor Syst. (MASS)*, Oct. 2017, pp. 530–534, doi: 10.1109/MASS.2017.88.

[42] B. Sumathi, "Grid search tuning of hyperparameters in random forest classifier for customer feedback sentiment prediction," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 9, pp. 173–178, 2020, doi: 10.14569/ijacsa.2020.0110920.

[43] R. Shu, T. Xia, L. Williams, and T. Menzies, "Better security bug report classification via hyperparameter optimization," 2019, *arXiv:1905.06872*.

[44] B. B. Yin and K. M. Liew, "Machine learning and materials informatics approaches for evaluating the interfacial properties of fiber-reinforced composites," *Compos. Struct.*, vol. 273, Oct. 2021, Art. no. 114328, doi: 10.1016/j.compstruct.2021.114328.

[45] D. N. Fente and D. Kumar Singh, "Weather forecasting using artificial neural network," in *Proc. 2nd Int. Conf. Inventive Commun. Comput. Technol. (ICICCT)*, Apr. 2018, pp. 1757–1761, doi: 10.1109/ICICCT.2018.8473167.

[46] T. Kim, B. Ha, and S. Hwangbo, "Online machine learning approach for system marginal price forecasting using multiple economic indicators: A novel model for real-time decision making," *Mach. Learn. Appl.*, vol. 14, Dec. 2023, Art. no. 100505.

[47] T. Takase, S. Oyama, and M. Kurihara, "Effective neural network training with adaptive learning rate based on training loss," *Neural Netw.*, vol. 101, pp. 68–78, May 2018, doi: 10.1016/j.neunet.2018.01.016.

[48] R. Chen, W. Zhang, and X. Wang, "Machine learning in tropical cyclone forecast modeling: A review," *Atmosphere*, vol. 11, no. 7, p. 676, Jun. 2020, doi: 10.3390/atmos11070676.

[49] M.-X. Wang, D. Huang, G. Wang, and D.-Q. Li, "SS-XGBoost: A machine learning framework for predicting newmark sliding displacements of slopes," *J. Geotech. Geoenvironmental Eng.*, vol. 146, no. 9, pp. 04020074-1–04020074-17, Sep. 2020, doi: 10.1061/(ASCE)GT.1943-5606.0002297.

[50] S. T. Ikram, A. K. Cherukuri, B. Poorva, P. S. Ushasree, Y. Zhang, X. Liu, and G. Li, "Anomaly detection using XGBoost ensemble of deep neural network models," *Cybern. Inf. Technol.*, vol. 21, no. 3, pp. 175–188, Sep. 2021, doi: 10.2478/cait-2021-0037.

[51] S. Cui, A. Sudjianto, A. Zhang, and R. Li, "Enhancing robustness of gradient-boosted decision trees through one-hot encoding and regularization," 2023, *arXiv:2304.13761*.

[52] G. Brassington, "Mean absolute error and root mean square error: Which is the better metric for assessing model performance?" *Geophys. Res. Abstr.*, vol. 19, p. 3574, Jan. 2017.

[53] H. Abbasimehr, M. Shabani, and M. Yousefi, "An optimized model using LSTM network for demand forecasting," *Comput. Ind. Eng.*, vol. 143, May 2020, Art. no. 106435, doi: 10.1016/j.cie.2020.106435.

[54] C. P. K. Basalirwa, "Delineation of Uganda into climatological rainfall zones using the method of principal component analysis," *Int. J. Climatol.*, vol. 15, no. 10, pp. 1161–1177, Oct. 1995, doi: 10.1002/joc.3370151008.

[55] W. M. Ridwan, M. Sapitang, A. Aziz, K. F. Kushiar, A. N. Ahmed, and A. El-Shafie, "Rainfall forecasting model using machine learning methods: Case study terengganu, Malaysia," *Ain Shams Eng. J.*, vol. 12, no. 2, pp. 1651–1663, Jun. 2021.

[56] E. Y. Boateng, J. Otoo, and D. A. Abaye, "Basic tenets of classification algorithms K-nearest-neighbor, support vector machine, random forest and neural network: A review," *J. Data Anal. Inf. Process.*, vol. 8, no. 4, pp. 341–357, 2020, doi: 10.4236/jdaip.2020.84020.

[57] M. Bataineh and T. Marler, "Neural network for regression problems with reduced training sets," *Neural Netw.*, vol. 95, pp. 1–9, Nov. 2017, doi: 10.1016/j.neunet.2017.07.018.

[58] S. O. Sulaiman, J. Shiri, H. Shiralizadeh, O. Kisi, and Z. M. Yaseen, "Precipitation pattern modeling using cross-station perception: Regional investigation," *Environ. Earth Sci.*, vol. 77, no. 19, pp. 1–11, Oct. 2018, doi: 10.1007/s12665-018-7898-0.

[59] M. T. Anwar, E. Winarno, W. Hadikurniawati, and M. Novita, "Rainfall prediction using extreme gradient boosting," *J. Phys., Conf. Ser.*, vol. 1869, no. 1, Apr. 2021, Art. no. 012078, doi: 10.1088/1742-6596/1869/1/012078.

[60] N. K. A. Appiah-Badu, Y. M. Missah, L. K. Amekudzi, N. Ussiph, T. Frimpong, and E. Ahene, "Rainfall prediction using machine learning algorithms for the various ecological zones of Ghana," *IEEE Access*, vol. 10, pp. 5069–5082, 2022, doi: 10.1109/ACCESS.2021.3139312.

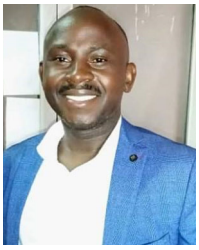
[61] K. Nongthombam, "Data analysis using Python," *Int. J. Eng. Res. Technol.*, vol. 10, no. 7, pp. 463–468, 2021.

[62] O. Kisi, S. M. Karimi, J. Shiri, and A. Keshavarzi, "Modelling long term monthly rainfall using geographical inputs: Assessing heuristic and geostatistical models," *Meteorological Appl.*, vol. 26, no. 4, pp. 698–710, Oct. 2019, doi: 10.1002/met.1797.

[63] S. Kareem, Z. J. Hamad, and S. Askar, "An evaluation of CNN and ANN in prediction weather forecasting: A review," *Sustain. Eng. Innov.*, vol. 3, no. 2, pp. 148–159, Oct. 2021, doi: 10.37868/sei.v3i2.id146.

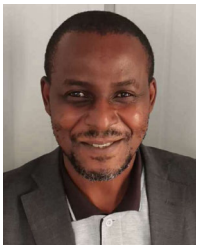
[64] P. Krammer, M. Kvassay, O. Habala, and L. Hluchý, "Short-term rainfall estimation by machine learning methods," in *Proc. IEEE 15th Int. Sci. Conf. Informat.*, Nov. 2019, pp. 89–94, doi: 10.1109/Informatics47936.2019.9119318.

[65] A. Lawal, S. Y. Yerima, D. O. Olago, P. O. Amingo, C. W. Kariuki, W. Wang'ombe, L. Olaka, L. Obiero, and S. O. Wandiga, "Evaluating machine learning models for rainfall prediction: A case study of Nyando in Kenya," in *Proc. IEEE 15th Int. Conf. Comput. Intell. Commun. Netw. (CICN)*, Kenya, Dec. 2023, pp. 264–271, doi: 10.1109/cicn59264.2023.10402269.
[66] G. Pattukandan Ganapathy, K. Srinivasan, D. Datta, C.-Y. Chang, O. Purohit, V. Zaalishvili, and O. Burdzieva, "Rainfall forecasting using machine learning algorithms for localized events," *Comput., Mater. Continua*, vol. 71, no. 3, pp. 6333–6350, 2022, doi: 10.32604/cmc.2022.023254.
[67] X. Lu, J. Fan, and J. Dong, "Forecasting multi-step ahead monthly reference evapotranspiration using hybrid extreme gradient boosting with GreyWolf optimization algorithm," *Comput. Model. Eng. Sci.*, vol. 125, no. 2, pp. 699–723, 2020, doi: 10.32604/cmes.2020.011004.



TUMUSIIME ANDREW GAHWERA received the bachelor's degree in information technology and the Master of Science degree in data communications and software engineering from Makerere University, Uganda, where he is currently pursuing the Ph.D. degree in information systems, under the supervision of Dr. Odongo Steven Eyobu and Dr. Mugume Isaac. Since 2010, he has been a Senior ICT Officer with the Ministry of Defense and Veteran Affairs (MODVA). Since 2012, he has been a part-time Assistant Lecturer and a Researcher with the Department of Information Systems, College of Computing and Information Science, Makerere University. His research interests include machine learning, deep transfer learning, and weather modeling.

ODONGO STEVEN EYOBU received the B.Sc. degree in computer science from Islamic University, Uganda, in 2004, the M.Sc. degree in data communication and software engineering from Makerere University, Uganda, in 2007, and the Ph.D. degree in electronics engineering from Kyungpook National University, South Korea, in 2018. He is currently a Lecturer with the School of Computing and Informatics Technology, Makerere University. His research interests include deep learning systems, indoor localization, vehicular communications, intelligent transportation systems, and wireless sensors.



MUGUME ISAAC received the B.Sc. degree in physics from Mbarara University of Sciences and Technology, in 2005, the M.Sc. degree in meteorology from Nanjing University of Information Sciences and Technology, in 2012, and the Ph.D. degree in meteorology from Makerere University, Uganda, in 2019. His specialization is numerical weather and climate prediction. He is currently the Director of forecasting services with Uganda National Meteorological Authority. He is also a Researcher at the College of Agricultural and Environmental Sciences, Makerere University.