Evaluating Machine Learning Models for Rainfall Prediction: A Case Study of Nyando in Kenya

Conference Paper · December 2023

CITATIONS
0

READS
80

9 authors, including:

Ahmad Lawal
De Montfort University
1 PUBLICATION **0** CITATIONS
SEE PROFILE

Suleiman Y. Yerima
British University in Dubai
73 PUBLICATIONS **2,398** CITATIONS
SEE PROFILE

Daniel Olago
University of Nairobi
159 PUBLICATIONS **3,974** CITATIONS
SEE PROFILE

Lydia Olaka
The Technical University of Kenya
57 PUBLICATIONS **1,843** CITATIONS
SEE PROFILE

Evaluating Machine Learning Models for Rainfall Prediction: A Case Study of Nyando in Kenya

Ahmad Lawal
Institute of Artificial Intelligence
De Montfort University
Leicester, United Kingdom
ahmad.lawal@dmu.ac.uk

Suleiman Y. Yerima
Cyber Technology Institute
De Montfort University
Leicester, United Kingdom
syerima@dmu.ac.uk

Daniel O. Olago
Institute for Climate Change & Adaptation
University of Nairobi
Nairobi, Kenya
dolago@uonbi.ac.ke

Philip Omondi Amingo
Department Climate Diagnostics and Predictions
IGAD Climate Prediction and Applications Centre
(ICPAC), Nairobi, Kenya
philip.omondi@gmail.com

Charles Wamagata Kariuki
Accounting and Finance Department
De Montfort University
Leicester, United Kingdom
charles.kariuki@dmu.ac.uk

Wangari Wang'ombe
Leicester Castle Business School
De Montfort University
Leicester, United Kingdom
wwangombe@dmu.ac.uk

Lydia Olaka
Dept. of Geoscience & the Environment
Technical University of Kenya
Nairobi, Kenya
lydiaolaka@tukenya.ac.ke

Linda Obiero
Dept. of Earth & Climate Sciences
University of Nairobi
Nairobi, Kenya
linda.obiero@uonbi.ac.ke

Shem Oyoo Wandiga
Managing Trustee
Centre for Science and Technology Innovation
P.O.Box 42792-00100, Nairobi, Kenya
shem.wandiga@csti.or.ke

***Abstract*—This paper presents a comprehensive evaluation of machine learning algorithms for rainfall prediction in the Nyando region. The study employs LSTM, XGBoost, Random Forest, and SVR algorithms, exploring both univariate and multivariate models to enhance the accuracy of predictions. Additionally, the paper examines three different outlier filtering methods and assesses their impact on the final prediction outcomes. The research endeavours to contribute valuable insights to the field of rainfall prediction and disaster management. By providing accurate and reliable rainfall predictions, this study aims to aid communities in the Nyando region and similar areas in their efforts to effectively mitigate the adverse impacts of extreme weather events.**

***Index Terms*—Rainfall prediction, Machine learning, LSTM, SVR, Random Forest, XGBoost, Disaster preparedness**

I. Introduction

Nyando, located in East Africa, is widely recognized for its susceptibility to extreme weather events, particularly droughts and floods. These events pose significant challenges to local communities, agricultural activities, and the region's socioeconomic stability. Hence, the development of a precise rainfall prediction model is imperative to facilitate proactive disaster readiness and effective mitigation of the adverse consequences arising from these occurrences.

The traditional approach to rainfall prediction has a historical reliance on statistical methods to establish correlations between rainfall and various meteorological factors, such as temperature, wind, pressure, and humidity, based on geographic coordinates [1]. However, rainfall dynamics' complex and non-linear nature presents inherent difficulties for accurate forecasting. Previous efforts have been made to address this non-linearity with techniques like Singular Spectrum Analysis, Empirical Mode Decomposition, and Wavelet analysis [2], [3]. Nevertheless, some of the mathematical and statistical models employed in such scenarios demand significant computing resources [4].

The task of predicting rainfall is inherently challenging due to its irregular patterns, which are further exacerbated by the influence of climate change. The unpredictability of rainfall events threatens communities and hinders their sustainable development [5]. Therefore, the accuracy of rainfall prediction is of paramount importance. Recent advancements in intelligent prediction models, encompassing various machine learning and deep learning approaches like Artificial Neural Networks, Random Forest, Support Vector Machine, XGBoost, LSTM, and GRU, have demonstrated promising outcomes in rainfall prediction across diverse regions [6]–[9]. Despite these strides, there remains potential for refining the accuracy of these models. Continuous research and innovation in machine learning techniques are pivotal for refining rainfall prediction methods and bolstering their reliability and efficacy.

This research is meticulously designed to evaluate a variety of machine learning models comprehensively applied to rainfall prediction in the Nyando region. The overarching objective is to comprehensively understand these models' strengths and limitations, thereby enhancing disaster preparedness and response strategies.

The study evaluates multivariate and univariate models for rainfall prediction and compares their performance. Multivari-

The 15th IEEE International Conference on Computational Intelligence and Communication Networks, CICN 2023, Bangkok Thailand , 22-23 Dec. 2023
(Accepted version)

ate models consider interrelationships among various meteorological variables, whereas univariate models solely rely on historical rainfall data. This comparative analysis aids in identifying the most appropriate approach for accurate rainfall prediction tailored to Nyando's unique conditions.

Furthermore, the research delves into the influence of outlier filtering techniques on model performance. This examination is of paramount importance in enhancing forecast reliability and refining disaster management practices.

By achieving these aims, this research extends invaluable insights into the domain of rainfall prediction and disaster management. The ultimate ambition is to provide tangible support to Nyando and similar regions in their pursuit of mitigating the adverse impacts of extreme weather events and fostering resilience against environmental challenges. Beyond its academic contributions, this study is poised to inform decision-making and disaster management practices directly.

The rest of the paper is structured as follows: Section II provides a review of related works in the field of rainfall prediction using machine learning models. Section III presents the dataset and methodology utilized in this study. The results and comparative analysis are discussed in Section IV, followed by the conclusions and future research directions in Section V.

II. Related Work

Accurate rainfall prediction plays a critical role in agriculture, water resource management, and disaster preparedness. Recent advancements in machine learning and deep learning have revolutionized rainfall prediction, offering novel approaches and enhanced accuracy. In this section, we provide an overview of research works that have harnessed the power of machine learning and deep learning techniques to predict rainfall in diverse geographical locations. The studies presented here shed light on the application of algorithms in the quest for more reliable and precise rainfall predictions.

Dawoodi et al. [10] conducted a comprehensive study in Maharashtra, India, covering the period from 2009 to 2018. The researchers compared the performance of two popular algorithms, Naive Bayes (NB) and Support Vector Machine (SVM), for rainfall prediction. Additionally, they explored the impact of different window size approaches, fixed and variable, on the prediction accuracy. Notably, SVM with variable window size achieved higher accuracy in rainfall prediction. Furthermore, the study identified meteorological factors such as atmospheric pressure, wind speed, and wind direction that significantly influenced the SVM model's decision-making and boundary determination, leading to more accurate rainfall predictions.

In Australia, Mahajan et al. [11] also examined NB and SVM and other machine learning algorithms for rainfall classification. The study considered algorithms such as NB, SVM, Random Forest (RF), and K-Nearest Neighbors (KNN) using the Australian weather dataset. However, for this study, the RF classifier exhibited the highest accuracy among the models tested, outperforming other methods in rainfall prediction tasks.

Liyew et al. [12] investigated the application of machine learning algorithms for daily rainfall prediction in Ethiopia. They examined Multiple Linear Regression (MLR), RF and XGBoost for their prediction using data spanning from 1999 to 2018. The study also focused on selecting relevant parameters for the prediction based on Pearson correlation values, retaining those above 0.2. When combined with selected environmental parameters, XGBoost emerged as the most suitable algorithm for daily rainfall prediction. However, the researchers suggested that future studies could further improve prediction accuracy by incorporating sensor and meteorological datasets and exploring big data analysis approaches.

Aguasca et al. [13] carried out a comparative study of established machine learning algorithms for monthly rainfall prediction classification in the Canary Islands, covering the period from 1976 to 2016. The study explored various machine learning techniques, including RF, Linear Discriminant Analysis (LDA), Generalized Linear Models (GLM), Support Vector Machines (SVM), Gradient Boosting (GB), XGBoost, and Logistic Model Trees (LMT). The researchers also investigated the influence of combining local meteorological variables and the North Atlantic Oscillation Index (NAO) with machine learning algorithms to improve predictive model accuracy. Among the models tested, XGBoost and GB demonstrated the highest accuracy, indicating the efficiency of machine learning in predicting rainfall in regions with complex orographic patterns. The study further highlighted the surprising finding that global variables like NAO had minimal influence, while local variables like Geopotential Height (GPH) played a more significant role in predictive models in complex orographic areas.

However, Baljon et al. [14] proposed a Function Fitting Artificial Neural Network classifier (FFANN) for rainfall classification in Saudi Arabia. The research aimed to forecast rainfall and assess its impact on crop yields using historical weather data from southern Saudi Arabia. In the preprocessing stage, the authors employed the Kalman filter to address missing or incorrect values, and they emphasized the significance of various normalizing approaches, including rescaling, standardization procedures, and rescaling to unit length. The proposed FFANN classifier demonstrated promising results and performed well in data classification, highlighting its potential utility in rainfall prediction applications.

In a study conducted in Lahore, Pakistan, with a dataset spanning from 2005 to 2017, Rahman et al. [15] proposed a novel approach for rainfall prediction using an integrated framework of multiple machine learning classifiers. The integrated classifiers include DT, NB, KNN and SVM. The uniqueness of this method lies in the utilization of fuzzy logic to construct a rule-based layer that refines the outputs from individual classifiers. By leveraging fuzzy logic, the framework achieves enhanced decision-making capabilities and improved overall performance in rainfall prediction.

Barrera et al. [5] compared machine learning and deep

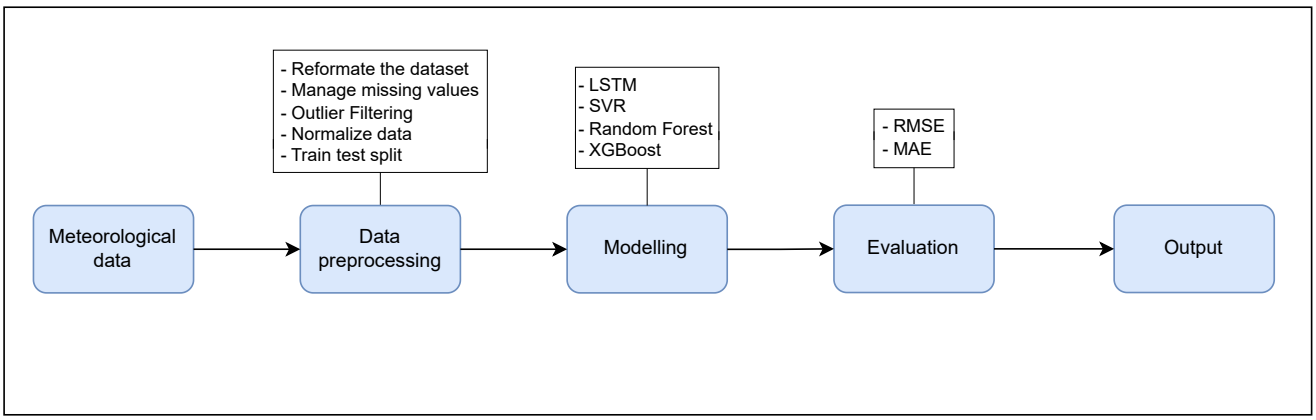

Fig. 1. Machine learning model

learning models for rainfall prediction in the United Kingdom using data spanning from 2000 to 2020. The authors proposed the utilization of advanced Long Short-Term Memory (LSTM) networks, including Bidirectional-LSTM and Stacked-LSTM models. These LSTM models were compared with other approaches, such as conventional LSTM, XGBoost, and an ensemble model. The results indicated that both Bidirectional-LSTM and Stacked-LSTM achieved high accuracy in rainfall prediction, with Stacked-LSTM slightly outperforming the others. However, it is worth noting that the Bidirectional-LSTM model inherited the limitation of LSTM-networks regarding generalization.

Kanchan et al. [16] conducted a detailed analysis of rainfall data for the Karnataka Subdivision in India, spanning from 1901 to 2017. The researchers employed three deep learning methods for rainfall prediction: Feedforward Neural Networks (FNN), Recurrent Neural Networks (RNN), and LSTM networks. Notably, the LSTM-optimized deep learning technique showed superior predictive outcomes with lower Mean Absolute Percentage Error (MAPE) and Root Mean Square Error (RMSE) values compared to the other models.

These studies collectively underscore the potential of machine learning techniques for precise rainfall prediction across diverse regions. The adaptability of machine learning in addressing key meteorological and hydrological challenges, such as disaster readiness, agricultural planning, flood prevention, and water resource management, is evident in these research efforts. Model effectiveness is influenced by factors like parameters, architecture, and training data characteristics. Comparing these models using Nyando's meteorological time series data could enhance applications targeting better quality of life and reduced socio-economic impacts of rainfall.

Such analyses aid extreme weather preparedness and decision-making to mitigate heavy rain's adverse effects on communities and infrastructure. This emphasizes tailoring machine learning models to specific regions, like Nyando, for accurate and localized rainfall predictions. Utilizing insights from these studies can improve model accuracy and reliability, contributing to superior water resource management, agricultural planning, and disaster readiness in Nyando.

III. Proposed Methodology

This section outlines the methodology used for rainfall prediction in Nyando, employing machine learning models. The study compares four models: LSTM, SVR, Random Forest, and XGBoost. The methodology begins with a description of the study area, Nyando, and continues with dataset collection, which includes hourly weather parameters. Tools used and data preprocessing steps, such as outlier handling and normalization, are then discussed. Subsequently, the machine learning models and performance metrics used for evaluation are presented.

A. Study Area

The study area, Nyando, is situated in western Kenya and is characterized by fertile plains, rolling hills, and numerous rivers, playing a pivotal role in agriculture and water resources. However, the area is susceptible to extreme weather events like droughts and floods, necessitating the need for accurate rainfall prediction.

B. Dataset Collection

For this study, the dataset was obtained from OpenWeatherMap [17], an online platform providing comprehensive weather data, including current, historical, and forecast data. The dataset comprises hourly records of various weather parameters, encompassing temperature, max temperature, min temperature, feels like, cloud percentage, humidity, pressure, rainfall, dew point, wind degree, wind gust, and wind speed. Their data sources encompass a range of inputs, including weather stations, radar, and satellite data. From 1979 to the present, this extensive temporal coverage enables a thorough examination of long-term weather patterns and trends in the Nyando.

C. Tools used

The study employed the free version of Google Colab [18], a cloud-based Jupyter notebook environment. This platform offers pre-installed libraries, GPU support, and collaborative

features to efficiently implement and compare LSTM, SVR, Random Forest, and XGBoost models for rainfall prediction in Nyando. It utilizes the Nvidia K80 GPU, equipped with 12 GB of memory, delivering a performance of 4.1 TFLOPS. Google Colab's cloud-based computing eliminates the need for local hardware resources, making it a cost-effective and accessible tool for machine learning research. The following models were used Python [19], Tensorflow [20], Keras [21], Scikit-learn [22]

D. Data Preprocessing

Data preprocessing is crucial for readying the dataset for machine learning models. The input variables encompass temperature, max temperature, min temperature, feels like temperature, cloud percentage, humidity, pressure, dew point, wind degree, wind gust, and wind speed. Rainfall is the targeted output variable. The process starts by handling missing values in the rainfall column, filling gaps with zeros for hours lacking rainfall.

Aggregation is performed to derive daily and monthly datasets, offering a broader time-based understanding of rainfall patterns. Normalization standardizes features, averting dominance of any specific attribute due to scale.

Two prediction approaches are explored: one incorporates outlier handling, while the other doesn't. This facilitates a comparative evaluation of outlier impact on predictive performance in rainfall prediction.

E. Outlier Handling

Outliers in the dataset can significantly affect the performance and accuracy of machine learning models. Three common outlier handling techniques were employed in this study to ensure robust and reliable rainfall predictions.

1) Interquartile Range (IQR) Method: The IQR method is used to identify outliers by measuring the spread of data within the middle 50% of the distribution. Data points that fall below $Q_1 - 1.5 \times IQR$ or above $Q_3 + 1.5 \times IQR$ are considered outliers, where Q_1 and Q_3 are the first and third quartiles, respectively. After carefully examining the dataset, we found that applying this method led to identifying extreme values as outliers. Instead of removing these outliers, we replaced them with capped values based on the calculated outlier thresholds. This approach allows us to retain the information on the extreme values while mitigating their potential impact on model training and ensuring a more robust analysis.

2) Mean and Standard Deviation Method: This technique involves identifying outliers based on their deviation from the mean. Data points that lie beyond 3 times the standard deviation from the mean are considered outliers. After analyzing the data, we found that using this method helped filter out data points that deviate significantly from the central tendency of the dataset. Rather than removing these outliers, we replaced them with capped values to maintain data integrity and contribute to the model's robustness during the analysis.

3) Mean Imputation : The approach used for outlier identification is straightforward, defining outliers as any data points with values less than or equal to the mean of the dataset. These identified outliers, which may correspond to abnormal conditions or measurement errors, are then replaced with the mean value. This process not only ensures the integrity of the data but also contributes to improving the accuracy of the model's predictions.

Before applying these outlier handling techniques, we thoroughly examined the dataset to understand the nature of the outliers and their potential impact on rainfall predictions in Nyando. The chosen outlier handling methods were then applied to the weather parameters to produce a refined dataset, with outliers replaced by capped values. By employing these outlier handling techniques, we aimed to enhance the quality of the input data and optimize the performance of the machine learning models for rainfall prediction in Nyando.

F. Machine Learning Models

This subsection outlines the different machine learning models employed for rainfall prediction in Nyando. Four distinct models, namely LSTM, SVR, Random Forest, and XGBoost, were selected for their ability to capture temporal patterns and handle complex relationships in time-series data. Section IV, presents the accuracy and reliability of each model based on their performance in rainfall prediction.

1) Long short-term memory (LSTM): LSTM, a variation of the recurrent neural network (RNN), excels in learning long-term relationships and retaining patterns from sequential and time-series data over prolonged periods [23]. It achieves this capability through the use of a gating mechanism that controls information flow within the network. The gating mechanism in LSTM consists of three components: an input gate, which manages new information added to the cell state; a forget gate, which controls the removal of information from the cell state; and an output gate, which governs the amount of information output from the cell state [24]. This sophisticated gating mechanism allows LSTM to effectively capture and process temporal dependencies, making it a powerful tool for accurate rainfall prediction in Nyando, where historical patterns significantly influence future weather occurrences.

2) Random Forest: Random Forest is a powerful supervised machine learning algorithm developed by Tin Kam Ho in 1995 [25]. It belongs to the category of ensemble learning techniques, where multiple decision trees are combined to form a robust predictive model. Each decision tree is constructed using constrained parameters and contributes to the final result.

By aggregating the predictions of multiple decision trees, Random Forest effectively reduces variance and enhances overall prediction accuracy. This makes it particularly suitable for handling high-dimensional and noisy data, common characteristics of rainfall prediction tasks in Nyando. The strength of Random Forest lies in its ability to capture complex relationships between weather parameters, enabling precise and reliable rainfall predictions.

3) Support Vector Regression (SVR): SVR, a powerful regression algorithm introduced by Vladimir Vapnik and colleagues in 1992 [26], is an extension of the well-known support vector machine (SVM) algorithm used extensively for classification tasks in supervised learning. In SVR, the main objective is to derive an estimating function from the actual outputs of the training data while ensuring a maximum deviation (H) from the true values. The key is to maintain the function as flat as possible, striking a balance between model complexity and accuracy [27].Unlike traditional regression machine learning algorithms that aim to minimize errors, SVM regression focuses on finding the best-fit line between the hyperplane and boundary line. This distinctive approach allows SVR to excel in both linear and nonlinear forecasting tasks, showcasing its versatility and efficacy.

4) Extreme Gradient Boosting (XGBoost): XGBoost is a cutting-edge and scalable tree boosting system based on the gradient boosting decision tree (GBDT) algorithm, introduced by Chen and his colleagues in 2016 [28]. Renowned for its remarkable speed, accuracy, and scalability, XGBoost operates as an ensemble learning algorithm, combining multiple decision trees to enhance prediction accuracy.

With its ability to handle both linear and nonlinear relationships, XGBoost is well-suited for capturing complex interactions between weather parameters in rainfall prediction. Its efficient implementation and impressive performance make it a valuable tool for precise and reliable forecasts in Nyando. By leveraging the strengths of ensemble methods and gradient boosting, XGBoost provides an effective solution for enhancing rainfall prediction accuracy.

G. Evaluation Metrics

To compare various machine learning model accuracy performances in predicting rainfall in Nyando, Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) were employed

1) Root Mean Squared Error (RMSE): RMSE is a widely used metric to quantify the differences between the predicted values and the actual observed values. It measures the average magnitude of errors, penalizing larger discrepancies between predictions and ground truth. A lower RMSE indicates that the model's predictions are closer to the actual values. Mathematically, RMSE is calculated as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}$$

Where:

- N is the total number of data samples.
- y_i is the actual observed rainfall value.
- $\hat{y}_i$ is the predicted rainfall value.

2) Mean Absolute Error (MAE): MAE is another performance metric that quantifies the absolute average difference between the predicted and actual values. Unlike RMSE, MAE does not penalize larger errors, providing a more intuitive measure of the average prediction error. A lower MAE indicates that the model's predictions are closer to the actual values. Mathematically, MAE is calculated as follows:

The Mean Absolute Error (MAE) is calculated as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i|$$

Where:

- N is the total number of data samples.
- y_i is the actual observed rainfall value.
- $\hat{y}_i$ is the predicted rainfall value.

H. Models' Parameters

The models were implemented using the parameters detailed in Table I. A validation period of two years was utilized, followed by a testing phase spanning another two years. For the monthly models, a sliding window of 12 months was employed, while the daily models operated on a sliding window of 14 days.

IV. Results and Discussion

In this section, we present the accuracy and reliability of each model based on their performance in rainfall prediction. This comprehensive evaluation will provide insights into selecting the most effective model for accurate forecasting and better disaster preparedness in Nyando.

A. Discussion

Accurate rainfall prediction is crucial for effective water resource management, agricultural planning, and disaster preparedness in Nyando. Timely and reliable forecasts enable authorities to take proactive measures in water allocation, irrigation planning, and flood control. Additionally, improved predictions can enhance early warning systems for droughts and floods, mitigating potential damages and loss of life.

1) Model Performance Comparison: The assessment of machine learning models for rainfall prediction in Nyando has provided valuable insights, offering a comprehensive understanding of their performance across varied scenarios. The subsequent table II and figures 2, 3 and accompanying paragraphs delve into the outcomes, focusing on different outlier treatment and data filtering methodologies.

XGBoost Multivariate consistently excelled in daily models, both with and without outlier handling. In contrast, Random Forest Multivariate performed least favourably without outlier treatment. LSTM Multivariate showed reduced accuracy, especially with IQR filtering, while SVR Univariate had the least accuracy using the STD and Mean Method.

For monthly models, XGBoost Multivariate demonstrated superior performance across all scenarios, lacking significant impact from outlier handling. However, LSTM Multivariate yielded the least favourable results without outlier treatment. LSTM Univariate exhibited relatively lower accuracy, particularly with IQR filtering. SVR Univariate fared poorly with the STD and Mean Method, and SVR Univariate's accuracy dropped when mean imputation was applied.

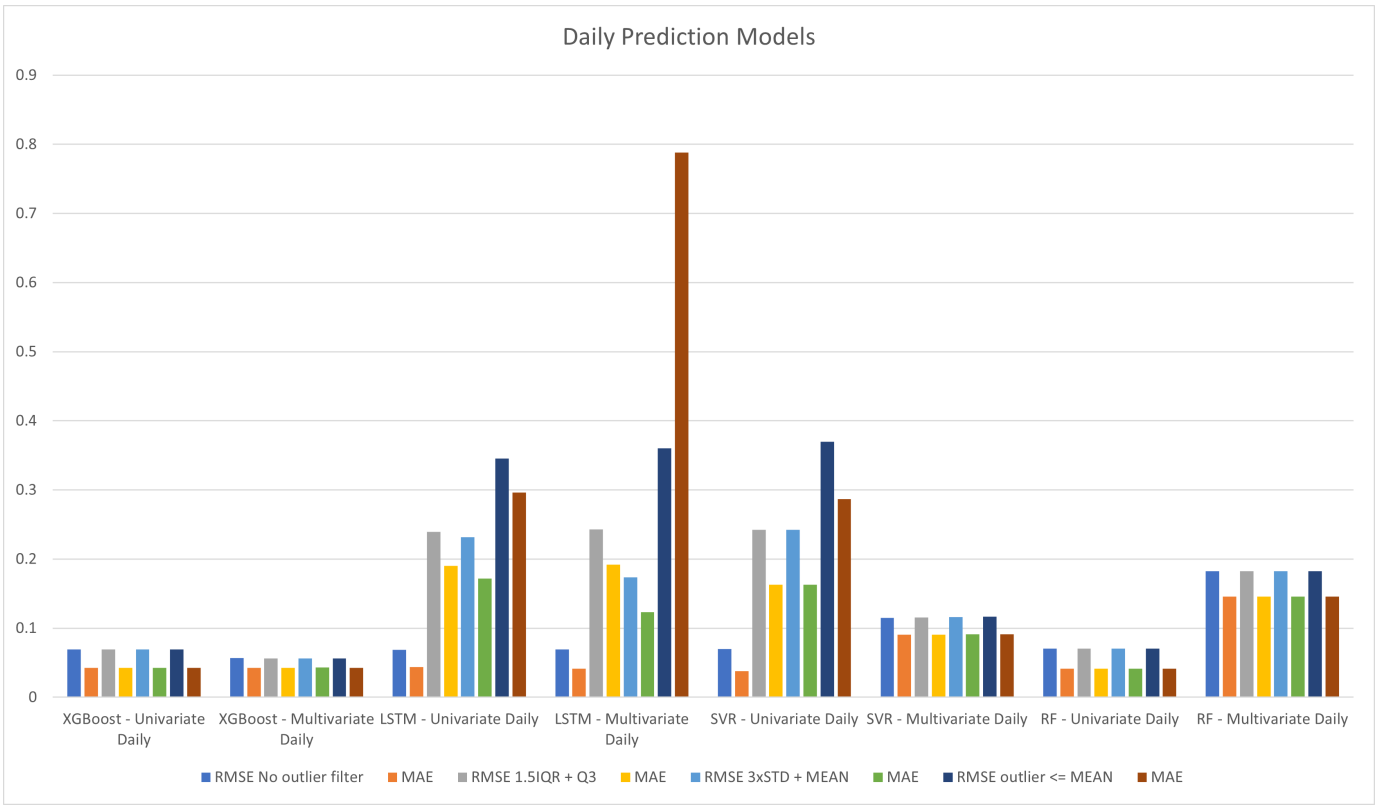

Fig. 2. Daily model performance

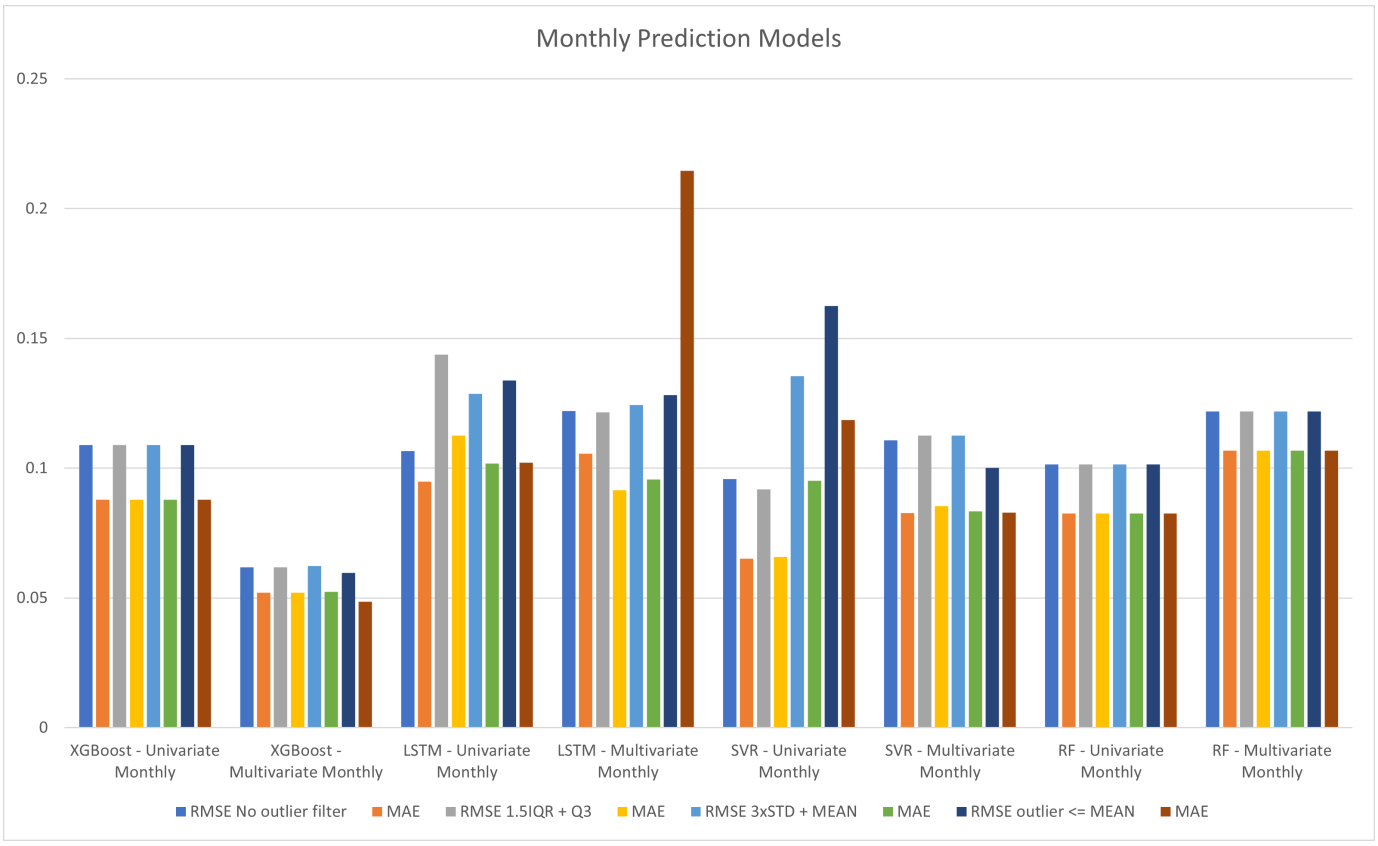

Fig. 3. Monthly model performance

TABLE I
MODEL'S PARAMETERS

Random Forest		XGBoost		SVR		LSTM	
n_estimators:	100	booster_type:	gblinear	kernel:	rbf	lstm_nodes:	128
max_depth:	None	learning_rate:	0.13	gamma:	1.2	dense_nodes:	16
min_samples_split	2	gamma:	6	C:	4	fct:	relu
min_samples_leaf:	1	tree_depth:	3	epsilon:	0.005	epochs:	50
bootstrap:	True	min_child_weight:	1			learning_rate:	0.001
random_state:	42	sample_size:	1			patience:	5

TABLE II
PERFORMANCE OF MACHINE LEARNING MODELS

Model	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
	No outlier filtering		IQR		Mean and STD		Mean Imputation	
XGBoost - Univariate Daily	0.06927	0.042424	0.069271	0.042424	0.069271	0.042424	0.069272	0.042421
XGBoost - Univariate Monthly	0.108966	0.08784	0.108966	0.08784	0.108967	0.08785	0.108962	0.087834
XGBoost - Multivariate Daily	0.05654	0.042529	0.056109	0.042638	0.056376	0.042886	0.056133	0.042698
XGBoost - Multivariate Monthly	0.061835	0.051988	0.061774	0.051969	0.062321	0.052378	0.059589	0.04855
LSTM - Univariate Daily	0.068348	0.043629	0.232485	0.173703	0.231758	0.171426	0.345467	0.295883
LSTM - Univariate Monthly	0.106572	0.09486	0.143641	0.112457	0.128649	0.101788	0.133799	0.102128
LSTM - Multivariate Daily	0.069228	0.041603	0.242715	0.192025	0.173514	0.123161	0.360181	0.788054
LSTM - Multivariate Monthly	0.12192	0.105571	0.121507	0.091489	0.124288	0.095692	0.128133	0.214525
SVR - Univariate Daily	0.069551	0.037723	0.242366	0.162822	0.242366	0.162822	0.369609	0.286431
SVR - Univariate Monthly	0.09581	0.065034	0.091863	0.06578	0.135461	0.095068	0.162461	0.118464
SVR - Multivariate Daily	0.115112	0.090385	0.115607	0.090771	0.115734	0.09103	0.116299	0.091139
SVR - Multivariate Monthly	0.110726	0.082649	0.112545	0.085392	0.112472	0.083427	0.100127	0.082834
RF - Univariate Daily	0.07041	0.04141	0.07041	0.04141	0.07041	0.04141	0.07041	0.04141
RF - Univariate Monthly	0.101458	0.082589	0.101458	0.082589	0.101458	0.082589	0.101458	0.082589
RF - Multivariate Daily	0.182297	0.145865	0.182297	0.145865	0.182297	0.145865	0.182297	0.145865
RF - Multivariate Monthly	0.121867	0.10666	0.121867	0.10666	0.121867	0.10666	0.121867	0.10666

2) Univariate vs Multivariate: This study's comparison of multivariate and univariate models illuminates their unique strengths. The experiment shows that multivariate models, adept at capturing intricate interdependencies, tend to provide a broader contextual understanding, potentially enhancing overall predictive accuracy. In contrast, while simpler, univariate models can still compete effectively by concentrating solely on the target variable. Opting for one approach over the other depends on factors such as dataset traits, feature interactions, and model complexity, unveiling the intricate nature of rainfall prediction modelling.

3) Effect of Outlier Treatment on Model Performance: XGBoost and Random Forest models were not affected by outlier treatment due to the robustness of boosting models. On the other hand, LSTM and SVR models experienced a decline in performance because of the loss of crucial information resulting from outlier removal or transformation.

4) Importance of Data Preprocessing: Data preprocessing is crucial for model performance. Substituting missing rainfall values with zeros and normalizing input features improve data analysis and model convergence. Outlier filtering did not significantly enhance model performance, but the preprocessing strategies elevated the model's reliability and efficacy in predicting rainfall.

V. CONCLUSIONS

This study on machine learning models for predicting rainfall in Nyando found that the XGBoost Multivariate model performed well in daily and monthly models, while LSTM models showed potential despite facing challenges in capturing long-term dependencies. Multivariate models tend to be more accurate than univariate models, as they can identify complex relationships. However, univariate models are simpler and may suffice for less complex problems. The choice depends on the complexity of the problem and the need for a more comprehensive analysis.

The model was significantly improved by using effective data preprocessing techniques, such as handling missing values and normalizing features. Although filtering outliers did not help, other preprocessing methods were successful. This research helps us understand the complexities of predicting rainfall and offers strategies to prevent disasters. It lays the groundwork for using machine learning to protect communities like Nyando and similar regions from rainfall-related disasters.

ACKNOWLEDGEMENTS

This study was supported by the ADRELO project (reference: EP/V004867/1) funded by EPSRC/ UKRI under Belmont Forum.

REFERENCES

[1] CL Wu and Kwok-Wing Chau. Prediction of rainfall time series using modular soft computingmethods. *Engineering applications of artificial intelligence*, 26(3):997–1007, 2013.

[2] Kai Gan, Shaolong Sun, Shouyang Wang, and Yunjie Wei. A secondary-decomposition-ensemble learning paradigm for forecasting pm2. 5 concentration. *Atmospheric Pollution Research*, 9(6):989–999, 2018.

[3] Yu Xiang, Ling Gou, Lihua He, Shoulu Xia, and Wenyong Wang. A svr–ann combined model based on ensemble emd for rainfall prediction. *Applied Soft Computing*, 73:874–883, 2018.

[4] Pritpal Singh and Bhogeswar Borah. Indian summer monsoon rainfall prediction using artificial neural network. *Stochastic environmental research and risk assessment*, 27:1585–1599, 2013.

[5] Ari Yair Barrera-Animas, Lukumon O Oyedele, Muhammad Bilal, Taofeek Dolapo Akinosho, Juan Manuel Davila Delgado, and Lukman Adewale Akanbi. Rainfall prediction: A comparative analysis of modern machine learning algorithms for time-series forecasting. *Machine Learning with Applications*, 7:100204, 2022.

[6] Jian Rong Ban, Qi Gou, and Ya Shi Li. Study on rainfall prediction of yibin city based on gru and xgboost. In *2022 4th International Conference on Advances in Computer Technology, Information Science and Communications (CTISC)*, pages 1–5. IEEE, 2022.

[7] MT Anwar, E Winarno, W Hadikurniawati, and M Novita. Rainfall prediction using extreme gradient boosting. In *Journal of Physics: Conference Series*, volume 1869, page 012078. IOP Publishing, 2021.

[8] Deepak Ranjan Nayak, Amitav Mahapatra, and Pranati Mishra. A survey on rainfall prediction using artificial neural network. *International journal of computer applications*, 72(16), 2013.

[9] Jiansheng Wu and Long Jin. Daily rainfall prediction with svr using a novel hybrid pso-sa algorithms. In *International Conference on High Performance Networking, Computing and Communication Systems*, pages 508–515. Springer, 2011.

[10] HH Dawoodi and MP Patil. Rainfall prediction for north maharashtra, india using advanced machine learning models. *Indian Journal of Science and Technology*, 16(13):956–966, 2023.

[11] Deepika Mahajan and Sandeep Sharma. Prediction of rainfall using machine learning. In *2022 Fourth International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT)*, pages 01–04. IEEE, 2022.

[12] Chalachew Muluken Liyew and Haileyesus Amsaya Melese. Machine learning techniques to predict daily rainfall amount. *Journal of Big Data*, 8:1–11, 2021.

[13] Ricardo Aguasca-Colomo, Dagoberto Castellanos-Nieves, and Máximo Méndez. Comparative analysis of rainfall prediction models using machine learning in islands with complex orography: Tenerife island. *Applied Sciences*, 9(22):4931, 2019.

[14] Mohammed Baljon and Sunil Kumar Sharma. Rainfall prediction rate in saudi arabia using improved machine learning techniques. *Water*, 15(4):826, 2023.

[15] Atta-ur Rahman, Sagheer Abbas, Mohammed Gollapalli, Rashad Ahmed, Shabib Aftab, Munir Ahmad, Muhammad Adnan Khan, and Amir Mosavi. Rainfall prediction system using machine learning fusion for smart cities. *Sensors*, 22(9), 2022.

[16] Pragati Kanchan and Nikhil Kumar Bakkappa Shardoor. Rainfall analysis and forecasting using deep learning technique. *Journal of Informatics Electrical and Electronics Engineering (JIEEE)*, 2(2):1–11, 2021.

[17] Openweathermap. https://openweathermap.org/.

[18] Google. Google colab. https://colab.research.google.com/.

[19] Guido Van Rossum and Fred L Drake Jr. *Python tutorial*, volume 620. Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands, 1995.

[20] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2015.

[21] François Chollet. Keras. *GitHub repository*, 2015.

[22] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[23] Ralf C Staudemeyer and Eric Rothstein Morris. Understanding lstm–a tutorial into long short-term memory recurrent neural networks. *arXiv preprint arXiv:1909.09586*, 2019.

[24] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[25] Tin Kam Ho. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE, 1995.

[26] Vladimir Vapnik, Isabel Guyon, and Trevor Hastie. Support vector machines. *Mach. Learn*, 20(3):273–297, 1995.

[27] Alex J Smola and Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and computing*, 14:199–222, 2004.

[28] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.