Global Journal of Engineering and Technology Advances
eISSN: 2582-5003
Cross Ref DOI: 10.30574/gjeta
Journal homepage: https://gjeta.com/

(RESEARCH ARTICLE)

Check for updates

Development of a machine learning model for precipitation forecasting in Kenya

Damaris Muthoki Mulinge [1,*], John M. Kihoro [2], Shadrack S. Madila [3] and Shem Mbandu Angolo [4]

[1] *Department of Information and Communication Technology, The Cooperative University of Kenya, Karen, Nairobi, Kenya.*
[2] *Department of Mathematical Sciences, The Cooperative University of Kenya, Karen, Nairobi, Kenya.*
[3] *Department of Information and Communication Technology, Moshi Cooperative University, Moshi, Tanzania.*
[4] *Department of Information and Communication Technology, The Cooperative University of Kenya, Karen, Nairobi, Kenya.*

Global Journal of Engineering and Technology Advances, 2025, 24(03), 043-050

Publication history: Received on 26 July 2025; revised on 29 August; accepted on 01 September 2025

Article DOI: https://doi.org/10.30574/gjeta.2025.24.3.0261

Abstract

Accurate precipitation forecasting is important for mitigating the impacts of climate variability in Kenya, where erratic rainfall events considerably affect agriculture, water control and disaster preparedness. Traditional methods such as ARIMA (Autoregressive Integrated Moving Average) and NWP (Numerical Weather Prediction) have shown to struggle with complex weather patterns due to linearity assumptions, high computational demands and limited spatial resolution. This research develops and evaluates an XGBoost-based machine learning model to enhance precipitation predictions both long-term and short-term. Utilizing a a 20-year weather dataset (2004 - 2024) with 7300 daily data records sourced from online Visual Crossing Weather Data, key features include temperature, humidity, wind speed, lagged precipitation (1-7), rolling means and seasonal encoding to capture bimodal rainfall patterns of the months of march-May, and October-December. Data processing involved min-max normalization of 0-1 range, feature selection, sin/cosine transformations for seasonal patterns and temperature-humidity interactions for connective modelling processes. The dataset used was split with 80% for training and 20% for testing and a temporal split ≤ 2020 for training and > 2020 for testing maintaining the chronological data order. The initial attempts exhibited poor performance with low R^2 = 0.066 and a high RMSE=1.06 hence leading to XGBoost binary classification shift to predict the likelihood of rain/no-rain tomorrow. Bayesian optimization and GridSearchCV hyperparameter tuning was applied with default 0.5 threshold adjustment for improved rain class sensitivity using classification metrics and resulted 76.76% accuracy, 70.14% precision, 33.36% recall, 45.12% F1-Score and ROC-AUC 0.75. Post-tuning accuracy by reducing the threshold to 0.3 to capture missed rainfall events: 73% accuracy, no-rain precision and recall 81%, 53% rain precision, 54% recall, F1-Score 54%. Temperature-humidity interaction as the top predictor in feature importance. The results contribute to improved precipitation prediction accuracy hence supporting decision making in agriculture, water resource management and early disaster preparedness in Kenya's climate vulnerable regions.

Keywords: Machine Learning; Climate variability; Precipitation forecasting; XGBoost; Binary Classification

1. Introduction

Accurate precipitation forecasting is critical for mitigating the impacts of climate change, especially in Kenya, which is vulnerable to extreme weather events. Many areas face challenges such as food insecurity and water scarcity due to unpredictable rainfall patterns. Precipitation variability poses a significant effect on the local communities dependent on natural resources, agricultural practices, and the region's socio-economic stability (IPCC, 2022) (KMD, 2023). Kenya's Climatic and weather conditions extremely contribute to food insecurity and water scarcity in about 75% of the country, and with erratic rainfall leading to droughts and floods that disrupt livelihoods (Affoh et al., 2022). Traditional approaches to weather forecasting though valuable, struggle with non-linear dynamics due to historical reliance on statistical correlations and have shown to struggle with non-linear dynamics, recent approaches such as Genetic

* Corresponding author: Damaris Muthoki Mulinge

Copyright © 2025 Author(s) retain the copyright of this article. This article is published under the terms of the Creative Commons Attribution Liscense 4.0.

algorithms and Neural Networks fail to capture the complex relationships between various factors that affect weather. Advanced Machine learning models particularly XGBoost, offer a data-driven approach to handle these challenges to enhance precipitation forecasting. This study develops and evaluates an XGBOOST for precipitation forecasting in Kenya focusing on temperature, humidity, wind speed, lagged precipitations (1-7), rolling means and seasonal encoding.

Forecasting precipitation plays a significant role in various sectors including agriculture, water resource management, and disaster preparedness. In Kenya, erratic rain patterns, prolonged droughts and floods intensify food insecurity, hinder economic development, and disrupt livelihoods (SAMWEL, 2021). Climate change impacts, water resources, agriculture, and ecosystems (Kogo et al., 2021). Traditional forecasting methods ARIMA and GCMs face data limitations, high computational costs, and low accuracy in localized forecasts, posing a need to data-driven methods. Kenya is located in East Africa. It is part of the Eastern region of the African continent, located below the Sahara Desert, defining the characteristics of sub-Saharan Africa. Most of its regions are susceptible to the effects of climate change which affect precipitation due to the mono-culture of rain-fed agriculture and minimal water resources (Pello et al., 2021) (Owino,2022). While traditional forecasting methods struggle with non-linear dynamics, XGBoost improves forecasting for preparedness. Climate vulnerability poses threat to agricultural productivity due to water scarcity, floods and drought. Currently traditional weather forecasting techniques such as Numerical Weather Prediction (NWP), synoptic forecasting fail to predict theshort and medium weather patterns within the regions accurately and understanding that the numerical weather prediction is dependent on extensive observational data that may be lacking, its substantial computational demands making real-time prediction a problem, and its coarse spatial resolution, which usually fails to accurately capture localized weather phenomena, highlighting a need for a Machine learning model, XGBoost.

- **The general objective:** To develop and evaluate XGBoost Machine Learning models to improve the accuracy of precipitation forecasting in Kenya.
- **Specific objectives**: i. Evaluate the XGBoost model using temperature, humidity, wind speed, and lagged precipitation as key variables. ii. Optimize via hyperparameter tuning to enhance precipitation forecasting accuracy. iii. Assess the implications in agriculture, water control management, and disaster preparedness.

1.1. Research Questions:

- How can XGBoost predict precipitation based on these variables?
- The performance of XGBoost in precipitation forecasting can be enhanced using

which optimization enhance XGBoost performance? iii. How can improved forecasts inform planning?

This research addresses traditional limitations in accuracy precipitation forecasting, promoting machine learning and supports crop rotation, irrigation practices and conservation measures to improve food security. Insights aid policy makers in strategy resilience with XGB Machine leaning model offering scalable solutions. The findings guide advanced machine learning models for short, medium and long-term forecasting's thus improving agricultural practices planning, water management (Sarma et al., 2024; Deo et al., 2022). The study the effectiveness of the application of Machine Learning, expanding environmental science applications in Kenya.

2. Materials and Methods

2.1. Data Collection

The study employs a quantitative data-driven paradigm using a 20 years historical weather data from visual crossing (2004 - 2024), with 7,300 daily observations of the key variables (temperature, humidity, windspeed, lagged precipitation, seasonal encoding and rolling means). Data was verified with meteorological reports to maintain consistency and completeness via systematic sampling to create datasets for real-time precipitation forecasting, improving data quality by removing biases.

Data Processing: Data cleaning to handle missing values was addresses using XGBoost (Aydin and Ozturk, 2021), with outlier detection using box plots. Min-max scaling normalized data to a 0-1 range (En-Nagre et al., 2024). Feature engineering included lags (1-7), sin/cosine transformations for bimodal rain variations for March to May and October to December and with temperature-humidity interactions for convection.

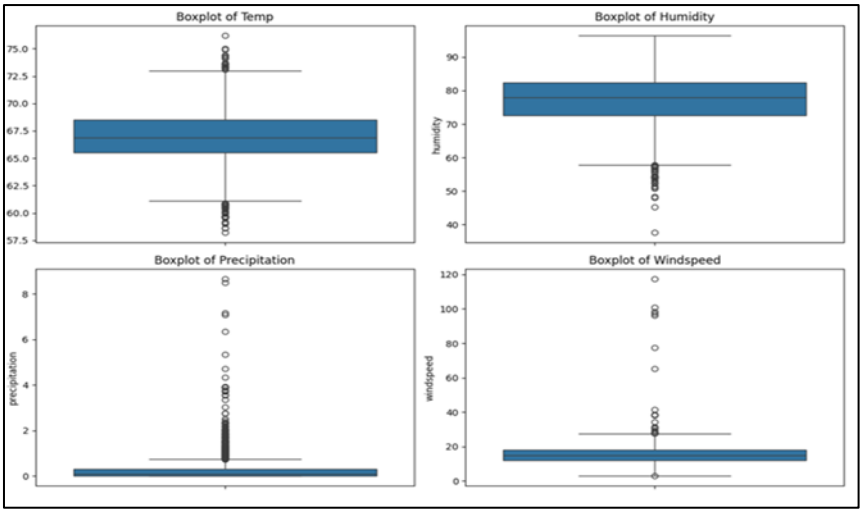


Figure 1 Outlier Detection

Time Series Analysis: Examining patterns over a time indicated inconsistencies particularly in precipitation(rainfall), with skewness right-tailed justifying XGBoost for capturing complex non-linear relationships.

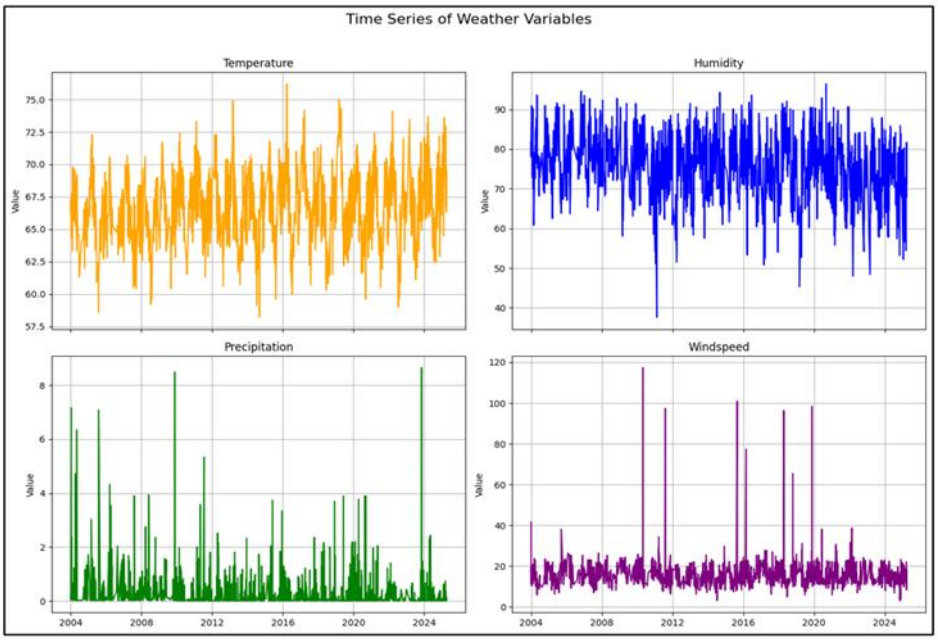


Figure 2 Time Series Analysis of Weather Variables

Descriptive Analysis: Average temperature was 66.94°C (SD=2.41. min 58.20, max 76.20), humidity 77.03% (SD= 7.56, min 37.60, max 96.40), precipitation 0.29mm, (SD=0.60), wind speed 15.68m/s (SD 6.13, MIN 2.90, max 117.40), highlighting precipitation irregularity.

Table 1 Descriptive Analysis: Summary Statistics

	Temperature	Humidity	Precipitation	Windspeed
count	2246.000000	2246.000000	2246.000000	2246.000000
mean	66.940338	77.030855	0.288359	15.675334
Std	2.409488	7.559530	0.603790	6.130008
min	58.200000	37.600000	0.004000	2.900000
25%	65.500000	72.600000	0.020000	12.100000
50%	66.900000	77.900000	0.079000	15.000000
75%	68.500000	82.400000	0.311000	18.200000
max	76.200000	96.400000	8.661000	117.400000

2.2. Model Development

The study employed a quantitative data-driven paradigm to forecast daily precipitation amounts using historical weather data, using regression-based predictions and to classify whether the trends result to rain or no-rain, using a temporal split of ≤ 2020 > 2020 for temporal realism. The development process included an orderly pipeline for accurate prediction using 80% training/20% testing with ≤ 2020train, > 2020 test, initializing, tuning, training and validation. At first the model was initialized for regression, then shifting binary classification if suboptimal using decision tress as base learners and with sequential error correction using a 20 years data from 200-2024. Hyperparameter tuning via GridSearchCV and Bayesian optimization for learning rate, max depth, estimators and subsample. Gradient Boosting trained XGBoost with weak learners added iteratively, k-fold cross-validation (k=5/10) for stability and prevent overfitting (Adnan et al, 2022). The temporal split avoided leakage and validation benchmarks against traditional models confirming generalization. Sensitivity analysis determined variable impact with regression metrics MSE, MAE, RMSE, R^2, and Classification metrics: accuracy, precision, recall, F1-Score and ROC-AUC.

Practical implications extend to agricultural practices for planting and irrigation, water management (allocation/storage), and disaster preparedness including alerts to minimize impacts.

2.3. Model Evaluation

Metrics included MAE, MSE, RMSE and R^2 for regression and accuracy, precision, recall, F1 and ROC-AUC for binary classification. Scikit-learn, pandas, numpy, matplotlib, and seaborn for analysis and visualization. Feature importance and ROC curves provided interpretability.

2.4. Ethical Considerations

Data handling ensured secure storage, transparency and fairness as per the Kenya Data Protection Act (2019). Systematic sampling reduced biases hence aligning with the principles of Machine Learning (ML) ethics for public interest.

3. Results

3.1. Model Performance

Initial XGB via regression approach exhibited poor results with low R^2 = 0.066, high RMSE = 1.06 leading to XGBoost Binary classification shift.

Table 2 Initial model performance via regression approach

R^2	0.06579746431625888
MSE	1.124557246672206
RMSE	1.06045143531998

3.2. Optimizing via Hyperparameter Tuning

GridSearchCV optimization using 300 n-estimators, 0.05 learning rate, max depth 5, 0.8 subsample.

Pre-tuning: Accuracy 76.76%, Precision 70.14%, 33.26% Recall, F1-score 45.12% and 0.75 ROC-AUC.

Tabel 3 Pre-tuned Model via Classification Approach

XGBoost Binary Classifier Performance	
Metric	Value
Accuracy	0.76
Precision	0.70
Recall	0.33
F1 Score	0.45
ROC AUC	0.74

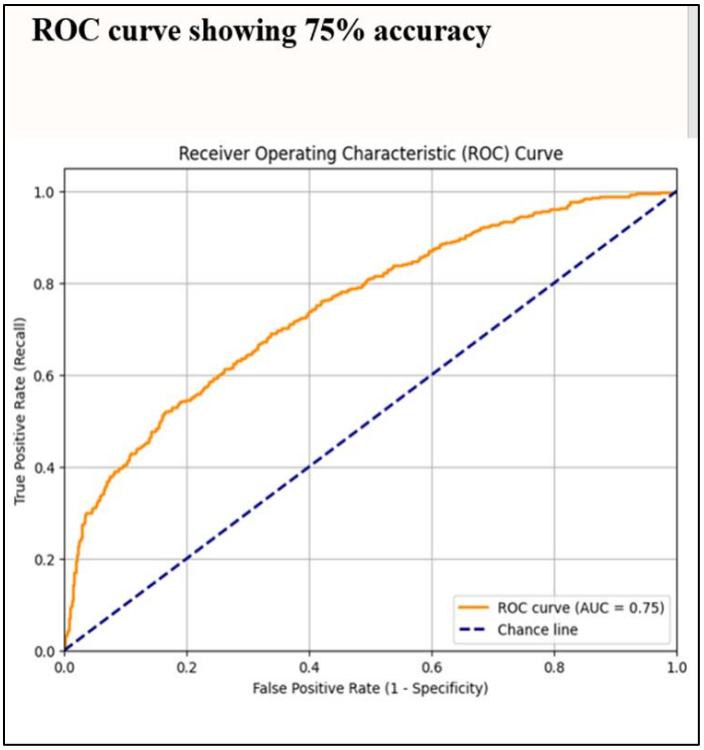


Figure 3 The ROC curve (AUC = 0.75)

Post-tuning: The threshold was reduced to 0.3 to capture missing rain events and it improved the precipitation prediction. Accuracy 73%, (reduced due to data imbalance), precision/recall/F1 for no-rain prediction 81%, rain precision 53%, and recall/precision for rain 54%. The Tuning improved the model's rain recall balance, precision, recall despite of a slight accuracy drop.

Table 4 Refined Model after further tuning

Metric	Class 0 (No-Rain)	Class 1 (Rain)	Macro avg	Weighted avg
Accuracy	-	-	-	0.73
Precision	0.81	0.53	0.67	0.73
Recall	0.81	0.54	0.67	0.73
F1 Score	0.81	0.54	0.67	0.73
Support	1104	445	1549	1549

3.3. Implications

The optimized model's accuracy of 0.73 and 0.54 recall for rain, the XGBoost Binary Classification model (the optimized machine learning model), supports the practical applications like in agriculture, accurate forecasts for rain/no-rain patterns for example the 80% precision for no rain help farmers in planting schedule planning especially during the Kenya's bimodal cycles for March to May and October to December rains. Predictions help in water reservoir efficiency for water resource management thus reducing the likelihood of floods because of reliable no-rain prediction. 0.54 Recall indicate fairly success in detecting rain events for further extreme occurrence improvements.

3.4. Feature Importance

A further analysis of the final XGBoost model was done to identify the top predictors of precipitation/rainfall. Temperature-humidity interaction reflected as the top precipitation predictor, humidity and temperature followed as the top features that influence precipitation prediction thus aligning with local weather patterns. However, poor predictors were the lag, sine/cosine variables which suggest that previous rain cycles definitely do not influence occurrences, as shown in the figure

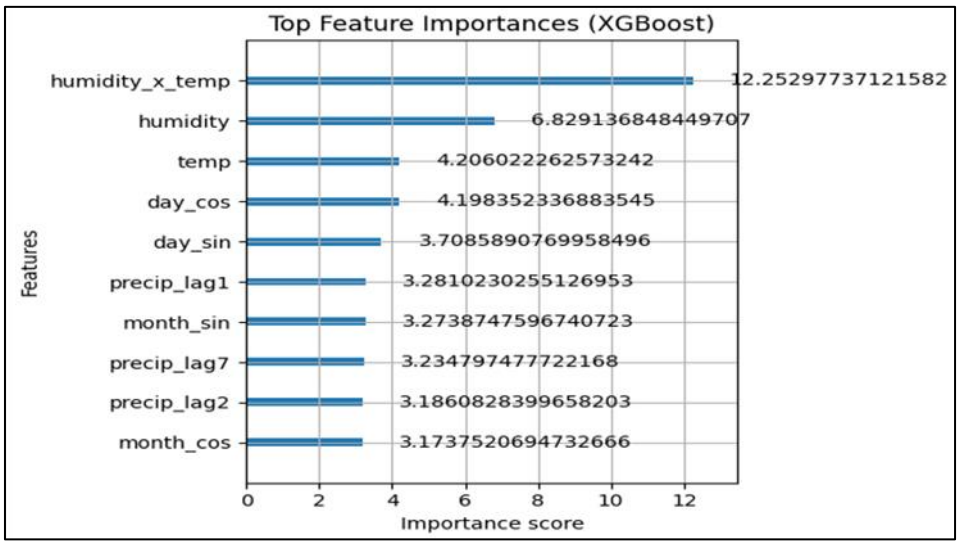


Figure 4 Feature Importance Results

4. Discussion

The discussion of these findings explores the level to which they be generalized. The initial XGBoost model via regression approach aimed on forecasting continuous precipitation amounts which showed poor performance with low R^2 of 0.066 and a high RMSE 1.06. Shifting to XGBoost Binary Classification model to forecast the likelihood of rain/no-rain reflected strong predictions more effectively. From the model it indicated that the optimized model learned to identify no rain and rain days by consistently distinguishing dry and wet conditions hence acquiring more stable capability when predicting rain cycles.

The study reflects the XGBoost maps meteorological predictors to binary precipitation outcomes effectively, with engineered features and temperature-humidity interactions confirming as the influential and reasonable drivers of precipitation predictors. The model's stability across data test shows the feature engineering robustness, temporal splits, and algorithmic design that align with regional literature of gradient boosting proving its superiority over simpler models on structured weather data. Although the XGBoost model reflected a high performance in identifying regular rainfall conditions, the models reduced sensitivity to rare extremes shows a typical trade-off seen in similar researches which can be reduced by stronger input sources like satellite, atmospheric pressure, using resampling and cost-sensitive techniques. Comparisons with East and West African studies validate the lagged precipitation values and humidity predictors as the suitability of the model to replicate with further fine-tuning or further adjusting.

The findings indicate the gradient boosting provides a well practical and interpretable method to long/short-term predictions for precipitations, while posing the need for richer data and customized approaches to capture unusual occurrences. The study contributes to the agreement that effective precipitation prediction depends on strong feature design and strong ensemble learning ensuring broader generalization with local validation.

Limitations

Even though the study achieved its objectives, providing meaningful insights, some limitations were acknowledged. The initial analysis relied on a single dataset, though detailed may not capture the current weather variability in broader conditions. The findings generalizability to other population may be limited and the study observational design indicate that normal relationships cannot be widely established and therefore the interpretation of the findings should be expressive than being broadly established. Oversampling of complex interactions by the binary with the data classification approach generally affected the results precision. Hyperparameter tuning sensitivity, data preprocessing choices, different parameter strategies might yield different results slightly. By identifying these constraints, future research should employ larger and more diverse datasets, consider other alternative modelling methods and include more variables to improve strong and the findings relevance.

5. Conclusion

This study evaluated the potential of using machine learning models to understand structure in a dataset to help guide decision making and operational improvement. The models discovered important relationships between variables and produced accurate predictions. Variability in the model's performance was be observed indicating the importance of model selection, tuning and validation as they can, in some cases, meaningfully impact results. Acceptable data quality and data structure (data preparation) were important to the findings of this study; indicating data preparation has a significant impact on results. The conclusions and recommendations of the study support the idea that there is value in creating and investing in, quality and higher quality datasets, hybrids practices, collaboration with subject matter experts, and regular model re-training for ongoing accuracy. In order to create the data and evidence needed when implementing workflows, around predictive analytics and maximizing their value, the model needs to be built into the workflow with reliable feedback cycles for ongoing accuracy.

Compliance with ethical standards

Acknowledgments

My deepest gratitude goes to God who provided all that was needed to complete this study. My sincere appreciation and honor to my supervisors Prof. John M Kihoro and Dr. Shadrack Stephen Madila for their remarkable contributions and constructive guidance, support, and invaluable feedback throughout this study that has been crucial to the success of this work, May the Almighty God reward them all. My appreciations to The Cooperative University of Kenya for provision of all necessary resources and a supportive environment for my study. Am thankful for the insightful and supportive friends and colleagues which has highly contributed to the success of this research. To my family members especially my mother Mrs. Agnes Henry who has been prayerful continually throughout my study, live long Mother. My special gratitude to my priest and spiritual head. Pastor Benjamin Mwirigi for prayers, encouragement and spiritual mentorship in all spheres of my life, more Grace. Proudly grateful to my benefactor, Dr. Per Hellsten for his immeasurable financial and academic support, throughout my academic journey, your generosity and belief in my potential gave me the strength and focus to persevere in my academic journey. To Ms. Lucy Maburi, your heartfelt support, your belief in my potential has been an inspiration throughout my studies. Above all i remain ever grateful to God for His fate, for indeed 'Destiny in God We trust'.

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] IPC (2022). Climate change 2022: Impacts, adaptation, and vulnerability. Working Group II contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press.

[2] KMD (2023). State of the climate in Kenya 2023. Technical report, Kenya Meteorological Department.

[3] Affoh, R., Zheng, H., Dangui, K., and Dissani, B. M. (2022). The impact of climate variability and change on food security in sub-saharan Africa: Perspective from panel data analysis. Sustainability, 14(2):759.

[4] SAMWEL, M. P. (2021). CLIMATE VARIABILITY AND FOOD SECURITY IN KISII COUNTY; KENYA. PhD thesis.

[5] Kogo, B. K., Kumar, L., and Koech, R. (2021). Climate change and variability in kenya: a review of impacts on agriculture and food security. Environment, development and sustainability, 23(1):23–43.

[6] Pello, K., Okinda, C., Liu, A., and Njagi, T. (2021). Factors affecting adaptation to climate change through agroforestry in kenya. Land, 10(4):371.

[7] Owino, D. O. (2022). Responding to impacts of climate change: A case study of kenya. Master's thesis, Oslo Metropolitan University.

[8] Deo, A., Karmakar, S., and Arora, A. (2022). Rainwater harvesting and water balance simulation-optimization scheme to plan sustainable second crop in small rain-fed systems. Journal of Environmental Management, 323:116135.

[9] Sarma, H. H., Paul, A., Kakoti, M., Talukdar, N., and Hazarika, P. (2024). Climate resilient agricultural strategies for enhanced sustainability and food security: A review. Plant Archives, 24(1):787–792.

[10] Aydin, Z. E. and Ozturk, Z. K. (2021). Performance analysis of xgboost classifier with missing data. In 1st Int. Conf. Comput. Mach. Intell.

[11] En-Nagre, K., Aqnouy, M., Ouarka, A., Naqvi, S. A. A., Bouizrou, I., El Messari, J. E. S., Tariq, A., Soufan, W., Li, W., and El-Askary, H. (2024). Assessment and prediction of meteorological drought using machine learning algorithms and climate data. Climate Risk Management, 45:100630.

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

[13] Kontopoulou, V. I., Panagopoulos, A. D., Kakkos, I., and Matsopoulos, G. K. (2023). A review of arima vs. machine learning approaches for time series forecasting in data-driven networks. Future Internet, 15(8):255.

[14] Papacharalampous, G., Tyralis, H., Doulamis, A., and Doulamis, N. (2023). Comparison of tree-based ensemble algorithms for merging satellite and earth-observed precipitation data at the daily time scale. Hydrology, 10(2):50.

[15] Noorbakhsh, M. (2022). Improving drought predictability in Africa by data-driven models. PhD thesis, University of Warwick.

[16] Habib-ur Rahman, M., Ahmad, A., Raza, A., Hasnain, M. U., Alharby, H. F., Alzahrani, Y. M., Bamagoos, A. A., Hakeem, K. R., Ahmad, S., Nasim, W., et al. (2022). Impact of climate change on agricultural production; issues, challenges, and opportunities in Asia. Frontiers in Plant Science, 13:925548.

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

[18] Adnan, M., Alarood, A. A. S., Uddin, M. I., and Rehman, I. (2022). Utilizing grid search cross-validation with adaptive boosting for augmenting performance of machine learning models. Peer Computer Science, 8:e803.

[19] Parmesan, C., Morecroft, M. D., and Trisurat, Y. (2022). Climate change 2022: Impacts, adaptation, and vulnerability. PhD thesis, GIEC

[20] Babu Nuthalapati, S., Nuthalapati, A., et al. (2024). Accurate weather forecasting with gradient boosting using machine learning. Int. J. Sci. Res. Arch, 12(2):408–422.