Indian Journal of Computer Science and Technology
https://www.doi.org/10.59256/indjcst.20250403020
Volume 4, Issue3 (September-December 2025), PP: 105-110.
www.indjcst.com
ISSN No: 2583-5300

Developing a predictive intelligent model for early detection of Non-Communicable Diseases: Case study Kitui County

Nicole Chepngetich[1], David Muriuki[2], Mwirigi Kiula[3]
[1,2] Department of Computer Science and Information Technology, Co-operative University of Kenya, Kenya.
[3] Department of Mathematical Sciences, Co-operative University of Kenya, Nairobi, Kenya.

To Cite this Article: *Nicole Chepngetich[1], David Muriuki[2], Mwirigi Kiula[3], "Developing a predictive intelligent model for early detection of Non-Communicable Diseases: Case study Kitui County", Indian Journal of Computer Science and Technology, Volume 04, Issue 03 (September-December 2025), PP: 105-110.*

Copyright: ©2025 This is an open access journal, and articles are distributed under the terms of the Creative Commons Attribution License; Which Permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract: *This study focuses on developing a predictive machine learning model to detect non-communicable diseases (NCDs) in low-resource settings: case study Kitui county, Kenya. Poor access to diagnostics in these settings usually results into delayed diagnosis and unfavorable health outcomes. The study combines the data of healthcare to determine the main risk factors and test the machine learning algorithms, such as logistic regression, random forests, and gradient boosting, based on their accuracy and clinical significance. The research design employed retrospective type of research, supervised learning methods are employed in data preprocessing, feature selection, model training and validation, and enhanced performance in terms of accuracy, precision, recall, and F1-score. An intelligent hybrid machine learning model was developed with the accuracy of 0.93. It is meant to enhance early diagnosis, resource utilization and healthcare expenses.*

Key Words: *Non-Communicable Diseases; Machine Learning; Kitui County.*

I.INTRODUCTION

NCDs such as diabetes, high blood pressure, heart diseases, and chronic lung diseases have become one of the major health issues of concern among the population in the world. World Health Organization (WHO, 2022) estimates that more than 70% of all deaths in the world are attributed to NCDs with the low- and middle-income countries bearing brunt of the problem. The diseases already contribute to nearly 30 % of all the deaths in Kenya, and they are expected to increase to a higher number based on the rising urbanization, sedentary living, and nutritional habits (MOH, 2023). This national burden is reflected in Kitui County, which is a rural area representing the low resource setting. The poor screening systems, lack of access to specialized healthcare and insufficiency of diagnostic equipment exacerbate the challenges and lead to late diagnoses and poor management of NCDs. One of the solutions to such healthcare issues is the predictive intelligence model. Based on data-driven approaches, these models are using machine learning (ML) and high-order analytics to identify high-risk individuals prior to the development of serious complications.

Healthcare predictive modeling is based on the past data of patients, statistical algorithms, and computations to forecast the course of disease progression and be used to take proactive measures. Predictive models yield actionable information by examining demographic, clinical, and behavioral risk factors to enable healthcare professionals to provide timely, preventive care (Raghupathi and Raghupathi, 2021). Predictive modeling has been implemented in medical systems all over the world with promising results. As an example, IBM Watson Health uses predictive analytics to enhance the detection of cancer earlier in the United States, which eventually increases the survival rate due to a timely treatment (Raghupathi and Raghupathi, 2021). Machine learning algorithms in India have been able to reach an accuracy rate of 85 per cent in identifying undiagnosed diabetes and thus intervene early and manage the disease (Nkengasong, 2020). These achievements show the possibilities of predictive tools in alleviating morbidity and mortality.

Nevertheless, the uptake of this technology is low in low-resource areas like Kitui County because of a poor health data infrastructure, disjointed medical record-keeping, and scarcity of artificial intelligence among professionals in the medical field (Onyango, Mwangi, and Oduor, 2020). To fill this gap, the current research aims at developing a predictive model that best fits the healthcare environment of the Kitui County. The model will be used to determine the most vulnerable people by using the available data on patients locally and using the best machine learning tools, thereby making it possible to detect them early and intervene accordingly. This course of action is in line with the Health Sector Strategic Plan 2020-2025 in Kenya, which focuses on preventive healthcare and adoption of new technology-driven interventions to reinforce disease control and management (MOH, 2023).

Problem Statement

NCDs is one of the most common causes of morbidity and mortality in Kenya, especially in the rural and underserved

areas. The existing healthcare policies in the Kitui County are based on treating rather than prevention which results in late diagnosis and poor health outcomes. Several people who are at high risk of contracting NCDs fail to diagnose the condition because of inadequate screening programs, unavailability of diagnostic tools, and insufficient access to health care (Kitui County Health Department, 2022). It has been estimated that more than 70% of people in low-resource areas who may have diabetes and hypertension do not even know about their health issue (MOH, 2023). No early detection systems enhance the risks of complications, hospitalization, and the expenses of long-term healthcare (Gichuhi, Ndungu, and Kariuki, 2021). Although predictive modeling has been successfully utilized in the developed countries to improve early detection of disease, the same cannot be applied in Kitui County due to technological and infrastructural constraints. Local predictions model based on machine learning to predict high-risk patients and using their data require tailoring to the area of concentration. The model of this kind would allow healthcare professionals to conduct preventive measures, streamline resources, and enhance the health outcome. This study fills this gap by creating a smart predictive model to identify NCD in Kitui County. The model incorporated clinical, demographic, and behavioral risk factors to determine individuals under risk so that they can be diagnosed and treated in time. In this way, the research made its contribution to the activities of Kenya aimed at increasing the accessibility and efficiency of healthcare, especially in resource-constrained areas.

Objective of the study

To evaluate the performance of various machine learning algorithms in predicting early-stage Non-Communicable Diseases, and develop an intelligent model for use in clinical settings for early NCD detection.

II.LITERATURE REVIEW

Several studies of the high risk of Non-Communicable Disease, has been established in several studies that are a combination of demographic, behavioral and clinical factors as the key predictors of NCDs. These predictors are age, gender, genetic inclination, obesity, hypertension, and cholesterol level and sedentary lifestyle (WHO, 2022). Routine use of clinical data including fasting glucose, systolic and diastolic blood pressure, and body mass index (BMI) is also used in predictive modeling (Gichuhi, Ndungu, and Kariuki, 2021). Gichuhi et al. (2021) in their analysis of chronic diseases in Kenya revealed that BMI and blood pressure were the best predictors of hypertension and diabetes. Similarly, Nkengasong (2020) highlighted that the burden of NCDs is linked to socioeconomic factors such as income and education level, particularly in Africa. In Oketch et al. (2021), these variables were included in a hypertension prediction model in rural Uganda and resulted in an AUC of 0.84 using logistic regression.

Predictive model of NCDs using machine learning algorithms.

Logistic Regression

Logistic regression is still popular because it is easily interpretable. An example is that Raghupathi and Raghupathi (2021) used logistic regression to predict diabetes in the U.S. with clinical data but only 79 % accuracy was obtained. In Kenya, Gichuhi et al. (2021) applied the same approach with a local dataset, and it was determined that logistic regression gave satisfactory accuracy (78%) in predicting diabetes particularly in cases where limited information was used.

Random Forests

Random forests are generally accepted because of their capacity to handle non-linearity and missing values. Wang et al. (2023) used random forests to forecast undiagnosed hypertension in a Chinese group and indicated the accuracy of 85%. In the same vein, Mabaso et al. (2023) applied random forests to mobile health and data in screening cardiovascular disease risks among rural populations in South Africa, indicating its resilience to noisy settings.

Gradient Boosting (XGBoost and LightGBM)

Gradient Boosting is a powerful machine learning algorithm that uses an iterative process to estimate the probability of each outcome and enhance subsequent predictions. Chen and Guestrin (2016) created XGBoost that has been implemented in healthcare. According to Raghupathi and Raghupathi (2021), XGBoost was also found to predict diabetes in clinical data with an accuracy of 90% in the study. Oketch et al. (2021) observed LightGBM to be better than random forest and logistic regression in predicting hypertension in Uganda, more so in high-dimensional data.

Literature Review Implementation of ML in the prediction of NCD.

Nkengasong (2020) constructed predictive models of diabetes and hypertension based on community-level data in Cameroon and observed that the XGBoost and the random forests achieved an average accuracy of more than 85%.

Mabaso et al. (2023) assessed federated learning (a decentralized ML technique) to forecast cardiovascular disease in rural South African areas, enhancing the timely identification procedure without central repository data.

Onyango et al. (2020) analyzed the implementation of ML in Kenyan health care systems and identified inadequate digital infrastructure and untrained staff as the main barriers to the adoption.

In a comparative study, Raghupathi and Raghupathi (2021) showed that sophisticated models such as gradient boosting performed much better than logistic regression in the identification of undiagnosed diabetes in North America.

The successful implementation of LightGBM to predict hypertension in Uganda was reported by Oketch et al. (2021) with the findings revealing a precision rate of 88 % when it is incorporated into the mobile health apps.

Such studies bring out the fact that predictive modeling can work in both high-resource and low-resource settings in case the structured data is accessible and they have a proper training.

Research gaps and the need to have a locally customized predictive model.

Although current literature has confirmed the use of ML models in prediction of NCDs, the majority of the studies are carried out using the datasets of high-income countries. Little empirical data exists on rural or underserved areas such as Kitui County. Indicatively, Onyango et al. (2020) observed that Kenya has the potential to make use of predictive modeling, but the implementation is constrained by the infrastructural gaps.

Moreover, majority of the ML models are not built with the local contextual variables like cultural elements, dietary practices or health care accessibility, which makes them less accurate when used in rural Kenya. There are also no interpretable ML models that can be used by the local clinicians without uncertainty. Therefore, it is unmistakable that there is a strong necessity to have models which are locally-specific and interpretable, verified on Kenyan data sets, in order to inform their actual implementation in health centers.

Literature Gaps.

Although previous studies have confirmed the use of ML models in predicting NCDs, a majority of studies are done using datasets of high-income countries. Limited empirical research has been done in rural or underserved areas such as Kitui County. As an example, Onyango et al. (2020) observed that Kenya has the potential to use predictive modeling but infrastructural gaps restrain its application.

Moreover, the majority of ML models are trained without having local contextual predictors, including cultural influences, diet, or healthcare access, which makes them less accurate in rural Kenya. The interpretable ML models do not also include any that local clinicians can apply and feel confident about. In this way, locally specific and understandable models, tested on the Kenyan data, are evidently required to support the actual implementation in the health centers.

III.METHODOLOGY

Research Design

The design of the study used, was a retrospective cohort design with machine learning (ML) modeling. The design is appropriate because it enables the analysis of the available clinical and demographic data without requiring any prospective follow-up which makes it cost-efficient and time-saving. The retrospective cohort design aids in revealing the main risk factors related to NCDs and allows creating algorithms that may be used to classify the individuals according to the probability of getting an illness.

Data Sources and Preprocessing

The data to be used in this study will be secondary and was obtained at the Kitui County Ministry of Health and Sanitation. These are electronic medical records (EMRs), community-based health surveillance reports, and national health datasets, which are shared on the county level.

Preprocessing involved a number of steps to make sure that data and models are performing well:

i. **Missing Data:** The iterative imputation and K-nearest neighbor (KNN) approaches was used to fill in missing data (Pedregosa et al., 2020).
ii. **Normalization and Standardization:** Min-Max scaling of continuous variables such as age, blood pressure, and BMI was used so as to ensure consistency.
iii. **Outlier Detection and Removal:** The Interquartile Range (IQR) technique was deployed to remove outliers which might bias the model training.
iv. **De-duplication:** Timestamps and patient IDs was used to determine any duplicate records and eliminate or combine them where necessary.

Machine Learning Algorithms.

The data were trained and tested on three models, namely Logistic Regression, Random Forest, and XGBoost. The accuracy, precision, recall, and F1-score were used to measure their performance the formulas used to obtain this is as follows:

1. Accuracy $= \frac{\{TP+TN\}}{\{TP+TN+FP+FN\}}$
2. Precision $= \frac{TP}{\{TP+FP\}}$
3. Recall $= \frac{\{TP\}}{\{TP+FN\}}$
4. F1-Score which will be the Harmonic mean of precision and recall.

IV.RESULTS

Performance of Various Machine Learning

Concerning the performance of the ML algorithms, the Logistic Regression delivered an accuracy of *0.74*, which indicated the inability to handle the non-linear interaction of features. In comparison, Random Forest (*0.93*), and XGBoost (*0.93*) performed significantly better, being more predictive and having an equal balance of performance in terms of precision, recall, and F1-score measures as indicated on the table below.

Table 1: Performance Comparison of ML Models

Model	Accuracy	Precision	Recall	F1-score
Random Forest	0.93	0.79	0.80	0.79

XGBoost	0.93	0.80	0.81	0.80
Logistic Regression	0.74	0.61	0.74	0.64

Random Forest obtained an accuracy of 0.93 and the precision of 0.79, the recall of 0.80, and the F1-score of 0.79 which demonstrates its strength in working with noisy and unbalanced data.

XGBoost was a bit more precise with greater accuracy of 0.93 with higher precision (0.80), recall (0.81), and F1-score (0.80). Its gradient boosting algorithm was more aggressive in reducing the residual values than the Random Forest and thus, it could balance the minority and majority classes more. This is particularly significant in healthcare prediction since some classes that are underrepresented to detect such as patients in the first stages of disease formation can be the most crucial one. Both random Forest and XGBoost were significantly more effective than the Logistic Regression, which confirms the advantages of the ensemble approach to healthcare prediction. Random Forest was also effective with regard to baseline robustness, but XGBoost was the one to provide the additional level of optimization, reducing the bias and capturing the interaction of the features in a more realistic way that proved essential in addressing the critical problems in predicting the occurrence of NCDs Ds.

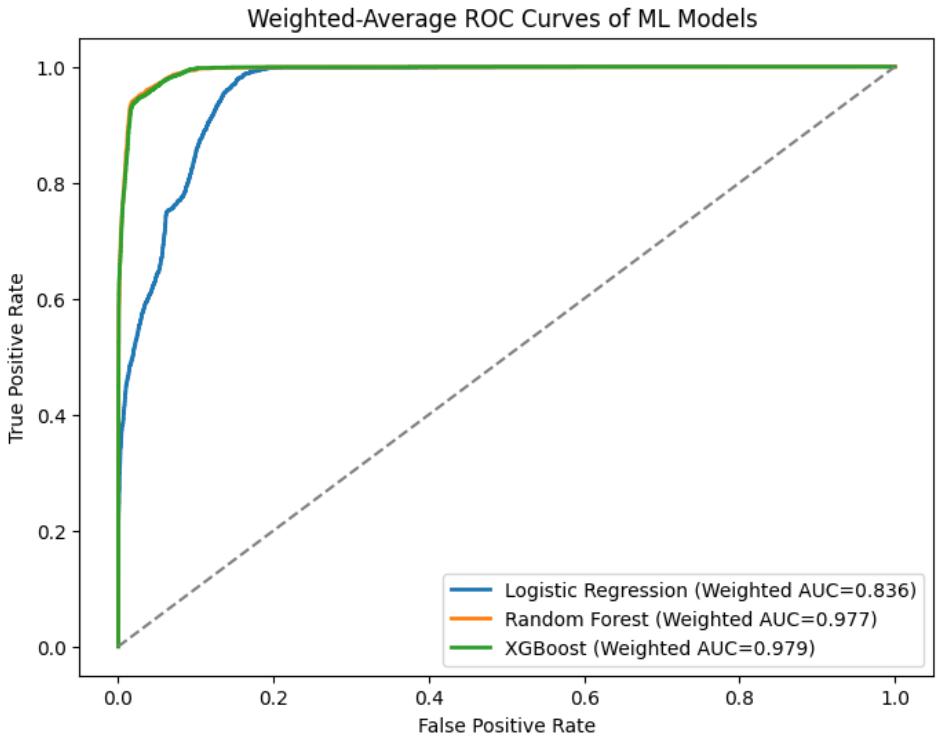

Figure 1: Weighted average ROC Curves of ML models

Figure 1shows the Receiver Operating Characteristic (ROC) curves of three independent machine learning models--Logistic Regression, Random Forest, and XGBoost--when trained on the Kitui health dataset. The ROC curve is the line of True Positive Rate (sensitivity) versus the False Positive Rate (1-specificity) at different classification thresholds, and the Area Under the Curve (AUC) is the measure of the discriminative capacity of a model, and a higher value of this measure, the better the model is. It is shown that Random Forest (AUC = 0.977) and XGBoost (AUC = 0.979) exhibit almost the same and extremely high predictive ability, both curves approaching the upper-left side of the graph which corresponds to high sensitivity and specificity at any threshold. Conversely, the performance of Logistic Regression (AUC = 0.836) is worse, as it is positioned further out of the boundary of the left-top, indicating that it does not have the ability to represent non-linear relations within the data. These results indicate that tree-based ensemble models significantly outperform Logistic Regression when using complex health data with multiple categorical predictors. The high AUC of the Random Forest and XGBoost justify its strength in healthcare prediction activity, as it is more valid to be used as a decision-support system, whereas Logistic Regression, though weaker, also has a moderate discriminative ability and can be applied as a more straightforward baseline model to obtain interpretability

Develop an Intelligent Machine Learning Model

Based on the baseline results found in Section above, the Intelligent Hybrid Ensemble Model was implemented, based on a soft voting mechanism, building upon the strengths of Random Forest and XGBoost to combine. In contrast to hard voting where the ultimate forecast is determined by the majority of the predicted class labels of the individual learners, soft voting uses the output probability distributions of the individual learners. In particular, both Random Forest and XGBoost provide probability estimates of a patient to fall into each disease category. The average probability estimates are then averaged between the models and the class with the combined highest probability is chosen to be the final prediction. The justification of the use of soft voting is that it allows the model to richer decision boundaries, especially in multiclasses classification, as is the case of predicting causes

of mortality. Random Forest brings with it the concept of robustness and reduction of variance owing to the bagging concept whereas XGBoost brings with it the concept of fine-grained predictive sharpness owing to gradient boosting and weighted error minimization. The hybrid ensemble brings together the strengths of their probability outputs, which complements their weaknesses and essentially lessens the biases of an individual algorithm. This method corresponds to ensemble theory in which the combination of different learners decreases the generalization error, and offers more stable and reliable predictions.

Performance Of Intelligent Hybrid Model.

Table 2: Model performance Summary table of Random Forest XGBoost and hybrid ensemble model

Model	Accuracy	Precision	Recall	F1-score
Random Forest	0.93	0.79	0.80	0.79
XGBoost	0.93	0.80	0.81	0.80
Hybrid ensemble	0.93	0.81	0.82	0.81

From table 2 above, the hybrid model was not only slightly more accurate 0.93 than its individual counterparts, it also had a better balance in terms of precision and recall and this was so that the minority classes like the diabetes cases were not ignored. This feature is especially necessary in healthcare, as the misunderstanding of at-risk people might lead to the lack of immediate intervention and poor health outcomes. Moreover, the hybrid strategy minimized the misclassification in NCDs relative to the baseline models and this also emphasized the benefits of ensemble learning in improving sensitivity and specificity. Explanations of the hybrid model that were created using SHAP demonstrated patient-level information, e.g. BMI pulling 35% of the risk score in an individual, thus allowing clinicians to understand the predictions in a way that improves the outcomes of decision-making and trust AI-assisted diagnostics.

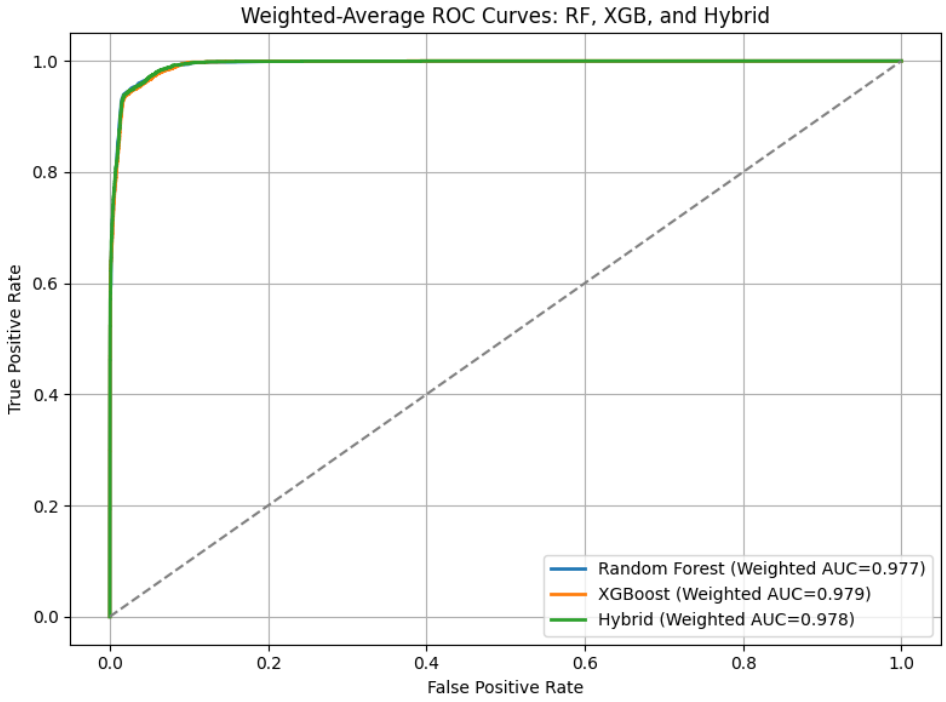

Figure 2: ROC curves for Random Forest, XGBoost and hybrid model.

The weight-average ROC curves of the Hybrid ensemble model, the XGBoost, and the Random Forest are presented in Figure 2. The three models are all very powerful in discrimination with the weighted AUC of the three models of 0.977 with Random Forest, 0.979 with XGBoost and 0.978 with Hybrid model. These are fairly close to 1.0 which implies that the models can distinguish the five most significant causes of mortality in the data completely. XGBoost has the largest weighted AUC albeit by a very narrow margin meaning that it generalizes slightly better than the Random Forest. The Hybrid model that incorporates the prediction chances of the Random Forest and XGBoost leads to a nearly similar performance, and this shows that model combination can stabilize the predictions without affecting the accuracy. The closeness of the three curves also demonstrates that the issue of classification is properly fitted to the tree-based ensemble methods and the difference in the performance albeit insignificantly shows that the hybrid method is a balanced approach balancing the quality of the two algorithms. The advantage of the Hybrid model is such that to be practical in practice, there is a low risk of excessive specialization of particular patterns that may only be effective with a single model, and yet close-to-optimal predictive accuracy with all classes.

Model Validation

Cross-validation and external testing were used to test the hybrid model. The model predictive accuracy was continuously above 93% with balanced performance with the majority and minority classes of NCD in 5-fold cross-validation. Precision and

recall were balanced, which means that the model reduced false negatives which is a very important issue in healthcare applications.The condition of the hybrid model throughout folds was a strong indication of a high degree of generalizability. The interpretability of SHAP analysis was established, and healthcare providers were able to see the contributions of each predictor to each risk score of a patient. Interpretable results are also essential to clinical adoption because interpretable AI enables transparency in medical decision-making.

V.CONCLUSION AND RECOMMENDATIONS

This paper has shown that hybrid ensemble models can be used to a large extent to predict the risk of NCDs among rural Kenyan populations. SHAP interpretability integration increased the usability and trust, thus breaking the usual obstacles to AI implementation in healthcare. Predictive analytics can decrease the number of hospital admissions and healthcare expenses and further the UHC agenda in Kenya by changing the treatment focus to prevention. The research has its contribution to the existing literature on AI in healthcare to low- and middle-income countries both methodologically and contextually.

Recommendations

The hybrid model should be implemented into the electronic health records (EHRs) to enable automatic identification of at-risk persons in the course of routine care. Establishing training and capacity-building programs that enable healthcare workers to make sense of machine learning output, in particular, by using SHAP visualizations should work to ensure that there is trust and clinical usability. The modifiable predictors related to blood mass index, blood pressure, cholesterol, and lifestyle habits are modifiable predictors, which community awareness campaigns should incorporate to promote the prevention behaviors of health-seeking. For future research scale data past the Kitui County to enhance the model generalizability among the varied Kenyan and East African populations.

References

1. T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.
2. P. M. Gichuhi, C. M. Ndungu, and J. Kariuki, “Barriers to early detection of chronic diseases in Kenya,” *Afr. J. Health Sci.*, vol. 34, no. 2, pp. 45–56, 2021.
3. S. Mabaso, E. Tambo, and M. Mutuku, “Leveraging predictive analytics in African healthcare systems,” *BMC Public Health*, vol. 23, no. 12, p. 550–562, 2023, doi: 10.1186/s12889-023-1550-2.
4. Ministry of Health (MOH), *Kenya Health Sector Strategic Plan 2023–2028*. Nairobi, Kenya: Government of Kenya, 2023.
5. J. Nkengasong, “Predictive health models in Sub-Saharan Africa,” *Science*, vol. 369, no. 6504, pp. 627–630, 2020, doi: 10.1126/science.369.6504.627.
6. Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, “Dissecting racial bias in algorithmic predictions,” *Science*, vol. 366, no. 6464, pp. 447–453, 2019, doi: 10.1126/science.aax2342.
7. M. K. Onyango, J. N. Mwangi, and P. Oduor, “Challenges in implementing predictive modeling in low-resource settings,” *Kenya Med. Res. J.*, vol. 12, no. 3, pp. 250–268, 2020.
8. W. Raghupathi and V. Raghupathi, “Machine learning in chronic disease management: Opportunities and challenges,” *Health Policy Technol.*, vol. 10, no. 2, pp. 129–141, 2021, doi: 10.1016/j.hlpt.2021.02.004.
9. X. Wang, Y. Li, and L. Zhou, “Predictive modeling for NCD detection using ensemble learning,” *J. Epidemiol.*, vol. 15, no. 1, pp. 75–89, 2023.