

Leveraging Machine Learning for Early Detection of Heart Disease in Akwa Ibom State, Nigeria

Samuel Olumide Adebayo and Chidera Nkechi Eze

Department of Computer-Aided Design and Engineering, Institute of Information Technology and Computer Science, Akwa Ibom University, Uyo, Nigeria

DOI: 10.5281/zenodo.14892307

Abstract

In middle-income countries as well as advanced economies of the world, one of the leading health problems is heart disease. Heart disease, otherwise known as cardiovascular disease, is a category of diseases and disorders usually characterized by abnormalities of the heart and the blood vessels, including coronary artery disease, heart failure, arrhythmias, and congenital heart defects. This study is very important as it aims to fill a knowledge gap and tackle the special challenges that primary healthcare systems in areas with few resources face leading to improved health results and more effective healthcare resource usage. In this study, we attempt to assess the efficacy of various machine learning (ML) methods for early heart diseases detection in Primary healthcare. We tested some ML algorithms – logistic regression, random forest, support vector machines (SVMs), K-nearest neighbors, CatBoost and XGBoost (XGB) using medical records from 2022 Primary Healthcare Facilities in Southern Nigeria. Each model showed different strengths in accuracy, precision, recall, and F1 score with logistic regression model achieving an overall accuracy of 85.61%. The Synthetic Minority OverSampling Technique (SMOTE) enabled us to mitigate class imbalance which boosted recall from 0.04 to 0.6016 and also balanced the F1 score from 0.08 to 0.3356; thus accurately identifies heart disease cases while maintaining fewer false negatives. Age (0.570193), daily smokes (0.370494), and blood pressure (0.365896) topped the list of heart risk factors. Blood sugar (0.189080), heart rate (0.096976), BMI (0.052322), and cholesterol (0.047680) also play a part in predicting overall risk. We recommend adding ML tools into routine healthcare, supported by policies, community outreach, targeted interventions, and continuous research to manage heart disease worldwide.

Keywords: Heart Disease, Machine Learning, Primary Healthcare, Feature Importance, Africa.

1.0 INTRODUCTION

Heart disease is a serious public health concern worldwide, causing substantial morbidity and mortality. The WHO overviews that heart diseases or cardiovascular diseases are the most common causes of death worldwide, accounting for approximately 17.9 million lives lost annually. These are blood vessel diseases and those of the heart that embrace coronary artery disease, heart failure, arrhythmias, and heart defects at birth. This is particularly true in low- and middle-income countries with meager health resources, whereas the principal risk factors, such as hypertension, diabetes, obesity, and smoking, are very prevalent and rising. The early detection capacity in Primary Health Care (PHC) systems is weak in large parts of the world, and this is important in ensuring auster

in the management and treatment of heart disease. Moreover, the late stage of detection mostly presents advanced stages of the disease, which are more complicated and expensive to treat, thus leading to worse health outcomes. In Africa, particularly in Nigeria, these challenges are further compounded by a lack of diagnostic capacity and specialized medical personnel in PHC settings.

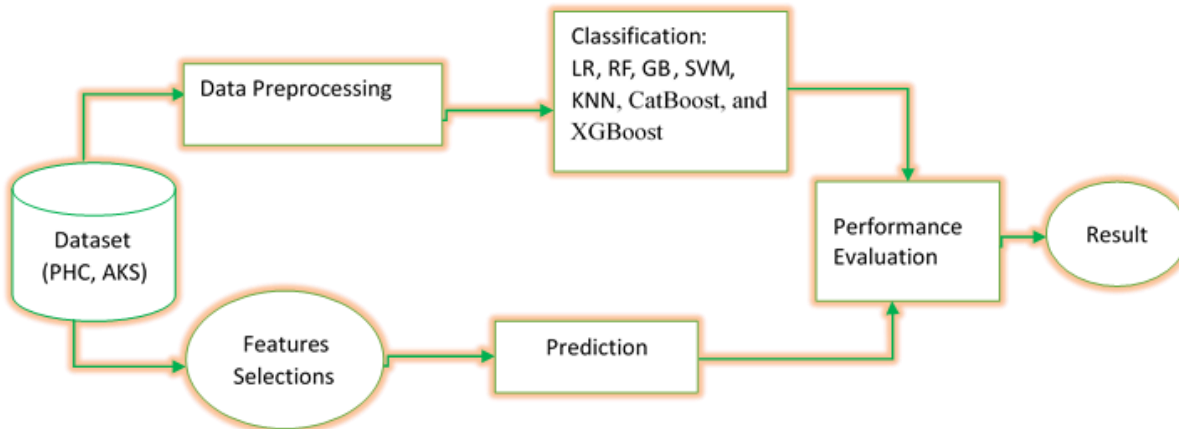
Modern research brings out an increasing trend in the application of machine learning for disease detection and management. For instance, in Nigeria, Ekle et al. (2023), tested the application of ML algorithms in predicting cardiovascular risk using the prevalent risk factors. In another related development, Samuel et al. (2024), applied ML methods for predicting under-five mortality in Nigeria, which gives credence to its versatility in handling sundry health challenges. A study conducted by Adeoye et al. (2020) in West Africa showed that ML models were very effective in identifying patients with high risks of hypertension and diabetes, which are some of the risk factors leading to heart disease. In addition, Kebede Kassaw et al. (2023) have elaborated on the application of ML techniques in establishing the predictors of anemia among under-five children in Ethiopia, further stressing the applicability of ML in health diagnostics across Africa. Indeed, pioneering works by Miotto et al. (2018) and Krittanawong et al. (2017) around the world have demonstrated superior performance of ML techniques over conventional statistical methods for cardiovascular event prediction, which holds huge potential to revolutionize healthcare systems.

Notwithstanding these developments, there are considerable gaps in the literature to the deployment of ML in resource-constrained settings such as Akwa Ibom State, Nigeria. Much of the literature focuses on high-resource environments, leaving a dearth of literature around how effective ML models will be and the very practical challenges of their application within Primary Health Care (PHC) settings in low-resource regions. Moreover, class imbalance in medical data sets can be one big issue that feeds biased models even more. For example, SMOTE was developed for solving the above-mentioned challenge: oversampling the minority class and improving predictive model performance (Chawla et al., 2002; He & Garcia, 2009). How these techniques work in the setting of heart disease detection using low-resource setting data needs further research.

Akwa Ibom State is a catchment area that is located in the southern part of Nigeria, with diverse healthcare needs and challenges. The State Health System is usually constituted by Primary Health Care centers, which form the first point of contact for most persons under its jurisdiction. However, most of them are usually plagued by a lack of diagnostic capacity and a paucity of specialized medical personnel. This work, therefore, aims at assessing the performance of various Machine Learning (ML) algorithms applied in the early detection of heart disease within Primary Health Care settings in Akwa Ibom State, Nigeria. In this study, different ML algorithms for the prediction of heart disease will be evaluated, important features that predict the disease will be identified, and the feasibility of implementing ML-based diagnostic tools assessed. Evidence-based recommendations on improving

the early detection and management of heart diseases will be presented to healthcare practitioners in Akwa Ibom State and globally. This research is very important, as it will fill the knowledge gap and address the challenges put forward toward the primary health care systems in low-resource regions and lead to improved health outcomes with better efficiency of the resources used.

2.0 MATERIALS AND METHODS **Figure 1: Methodology workflow**



2.1 Study Area:

The study was conducted in Akwa Ibom State, Nigeria, with primary health care settings in view, located in the southern part of Nigeria, which is made up of thirty-one local government areas.

2.2 Sample Collection:

In this study, data was obtained from Medical Records Units, Primary Health Care, Akwa Ibom State, Nigeria which comprised an anonymous set of patients' records collected during a specific time frame. All participants in this study provided informed consent and ethical approval was obtained. The dataset comprised several features: 'Male' indicating gender (0 for female, 1 for male), 'Age' of the patient, 'Education' level, 'CurrentSmoker' indicating if the patient currently smokes (1 for yes, 0 for no), 'CigsPerDay' showing the number of cigarettes smoked per day if the patient is a smoker, 'BPMeds' indicating if the patient is on blood pressure medications (1 for yes, 0 for no), 'prevetentStroke' indicating if the patient has a history of stroke (1 for yes, 0 for no), 'prevalentHyp' indicating if the patient has prevalent hypertension (1 for yes, 0 for no), 'diabetes' indicating if the patient has diabetes (1 for yes, 0 for no), 'totChol' representing the total cholesterol level, 'sysBP' and 'diaBP' representing the systolic and diastolic blood pressure respectively, 'BMI' representing the Body Mass Index, 'heartRate' representing the heart rate, 'Glucose' representing the glucose level, and

‘TenYearCHD’ representing the ten-year risk of coronary heart disease. These features collectively provided a comprehensive overview of the health status of the patients involved in the study.

2.3 Experimental platform:

In this work, we implemented our machine learning models using Google Colab due to its cloud-based environment that provides free-of-cost access to considerable computational resources. These resources include a virtual machine powered by an NVIDIA Tesla K80 GPU, up to 52 GB of RAM, and up to 100 GB of hard disk space, critical in the handling of large datasets and efficiently training complex models. The interface is much more familiar, comfortable, and easy to use in Google Colab, a much more recognizable Jupyter Notebook for writing and running code on the fly. Nearly all of the important libraries involving data science tasks within this platform are preinstalled, which includes TensorFlow, PyTorch, and Scikit-Learn. This thus enables us to set up and run our experiments quite fast. Stability and reliability in the operating system, particularly Ubuntu, further ensured consistent and robust for use in conducting research. These features made Google Colab the best option for developing and testing our machine learning models with respect to the early detection of heart disease in primary healthcare settings in Akwa Ibom, Nigeria.

2.4 Data Preprocessing:

Several steps were involved in data preprocessing. First, it comprised the initial exploration about getting to know the dataset and understanding any probable issues. Missing values were dealt with using techniques of imputation so there were no missing values in the dataset. Continuous features were normalized to be on a similar scale. SMOTE was employed in the case of class imbalance to rebalance the dataset.

2.5 Measurement of parameters:

This research included a wide array of demographic, lifestyle, and clinical parameters. The demographic data included information about the age and gender of participants, education level. Among lifestyle factors, smoking status and number of cigarettes per day were taken into consideration. Medical history parameters included prevalent hypertension, diabetes status, history of previous strokes, and existing heart diseases. Clinical measures for blood pressure in its systolic and diastolic state, fasting levels of total cholesterol plus glucose, body mass index, and heart rate were taken.

2.6 Exploratory Analysis

Exploratory data analysis has been done to narrow down the most important predictors for heart disease. This involved analyzing and ranking features with statistical methods and domain knowledge. We computed correlation coefficients to establish different extents by which variables relate to each other. The correlation computes the degree of the linear relationship between two variables ranging from -1 (perfect negative correlation) to 1 (perfect positive correlation), with 0 indicating no linear relationship. To identify meaningful

relationships, we focused on correlations from 0.30 and above. This threshold signifies moderate to strong relationships, suggesting variables that potentially influence the prediction of heart diseases. By examining the correlation plot, correlations from 0.30 were quickly identifiable, guiding us to prioritize variables such as age, cholesterol levels, blood pressure, etc., which exhibit notable associations with our target outcomes. Knowing these relationships will enable us to select for our predictive modeling task at hand only relevant features and be certain that our analysis is based on statistically significant relationships present in our dataset.

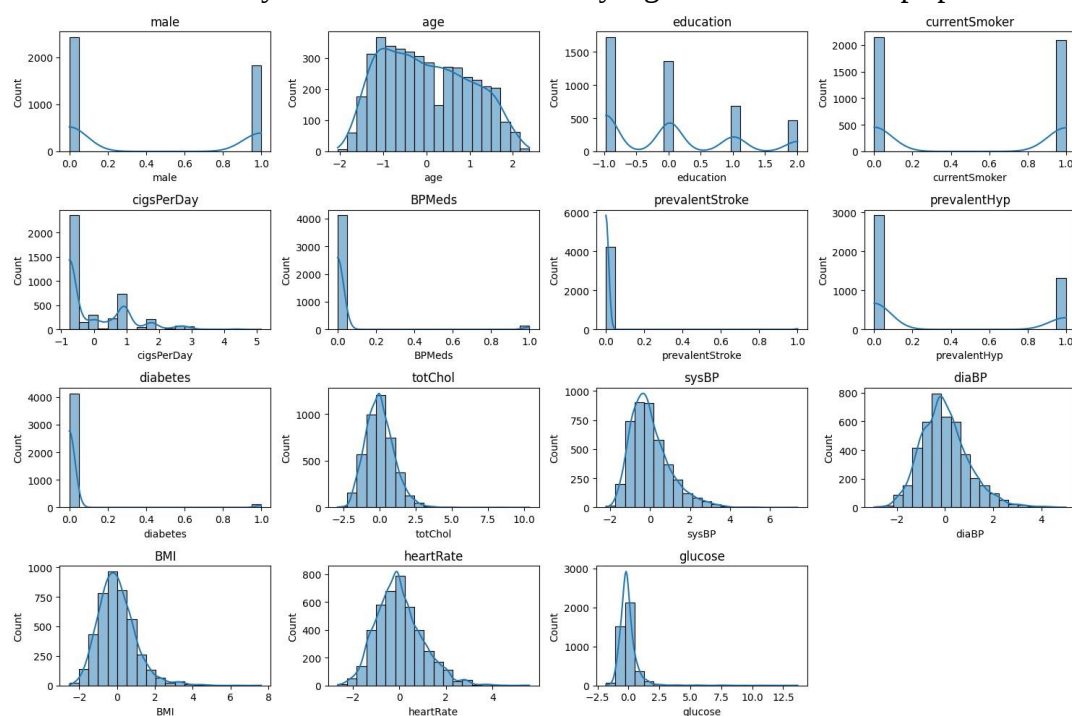


Figure 2: Distribution of Numerical Features

This figure shows the histograms of various numerical features in the dataset. These histograms highlight the distribution patterns of each feature, providing insights into their central tendencies, variances, and potential skewness.

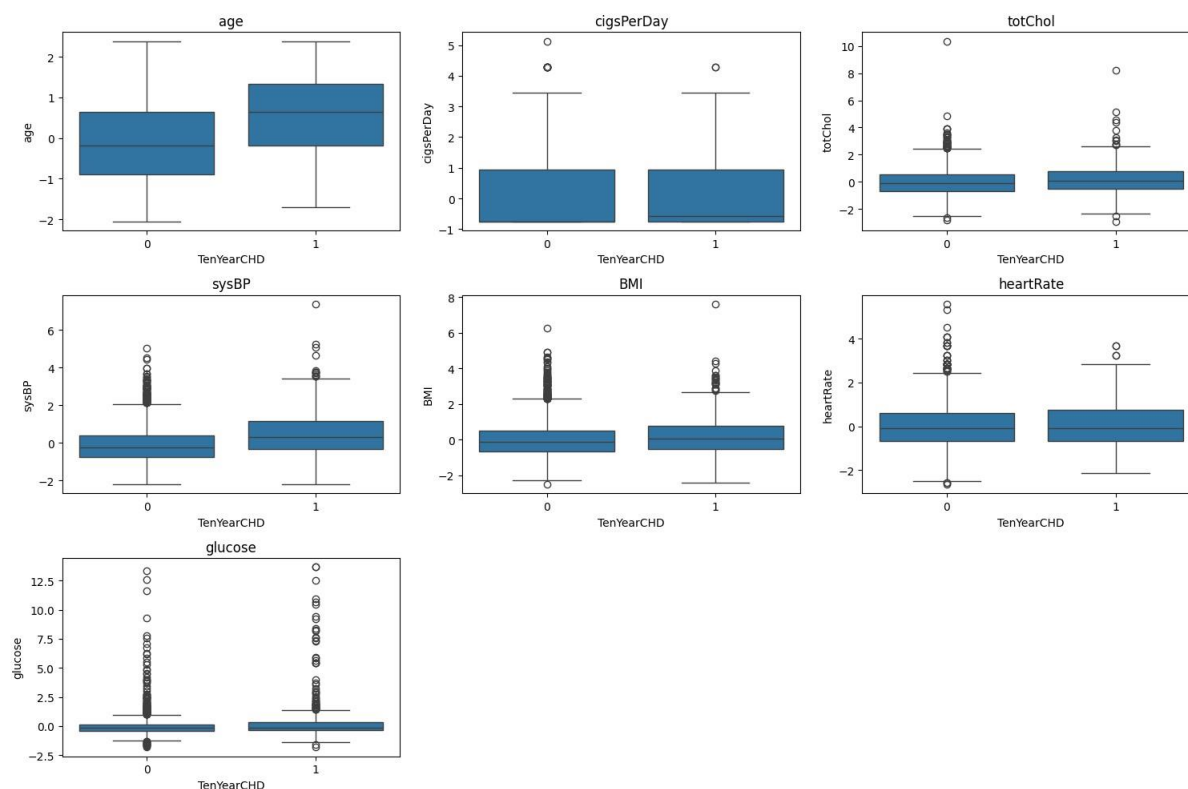


Figure 3: Boxplots of Selected Features by Target Variable

This figure presents boxplots illustrating the distribution of selected features (age, cigarettes per day, total cholesterol, systolic blood pressure, BMI, heart rate, and glucose) grouped by the target variable (TenYearCHD). These boxplots help identify the differences in feature distributions across the categories of the target variable, revealing potential predictors of heart disease.

Artificial Intelligence, Machine Learning, and Data Science Journal

Research Article

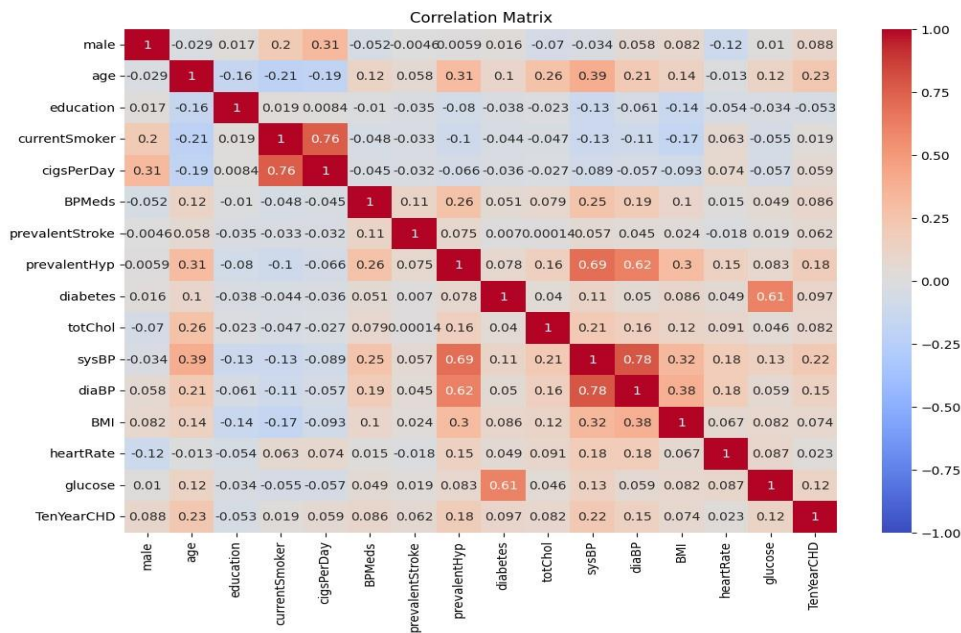


Figure 4:
Correlation Matrix for the Dataset

Figure 4 shows the correlation matrix for the dataset. The matrix displays pairwise correlation coefficients between features, with values ranging from -1 to 1. Positive values indicate a positive linear relationship, while negative values indicate a negative linear relationship. Correlations from 0.30 and above are emphasized as they represent moderate to strong relationships that are potentially significant for predicting heart disease. This visualization helps identify which features are strongly related to each other and to the target variable, aiding in feature selection and understanding of feature interactions.

2.7 Predictive models

Various algorithms of machine learning were adapted in the study to predict TenYearCHD, including Logistic Regression, Random Forest, Support Vector Machine, K-Nearest Neighbors, CatBoost, and XGBoost. **Logistic Regression:** Logistic regression is a linear model for classifying binary variables based on one or more predictor variables, which estimates the probability of a binary outcome. Here, the logistic function models the binary dependent variable. **Random Forest:** Random Forest is an ensemble learning method whereby many decision trees are constructed during training and classifies by outputting the mode of classes. It is robust to overfitting issues and amazingly performs well on varied datasets. **K-Nearest Neighbors (KNN):** KNN represents a very

simple class of instance-based learning algorithms that classifies the data point based on how its neighbors are classified. The algorithm is very intuitive and performs well with smaller datasets. **Support Vector Machine (SVM):** A powerful classification technique that encapsulates the model of the best fitting hyperplane, which separates classes. It is very effective against high dimensional spaces and evident margin separation. **CatBoost:** CatBoost is a gradient-boosting algorithm that processes categorical features on its own. Developed to be both fast and accurate, it often performs better than other boosting algorithms on categorical data. **XGBoost:** XGBoost is actually an efficient and scalable implementation of gradient boosting. In machine learning competitions, it is found at the top in most of them and is widely known for its performance and speed.

All the models were trained and tested using a stratified train-test split method so that both sets would still contain the same proportion of samples for each class of the target variable, TenYearCHD, to avoid bias and ensure generalizability.

2.8 Model Evaluation and Results:

The performance of each model was evaluated using several measures. These included **Accuracy**, which is the proportion of true results (both true positives and true negatives) among the total cases taken for evaluation. **Precision** was also calculated, defined as the fraction of true positives among all positive predictions. **Recall** was measured which is the ratio of correctly predicted positive observations to the all observations in actual class. Lastly, the **F1 Score** was computed, which is the harmonic mean of precision and recall. These measures collectively provided a comprehensive assessment of each model's performance.

Mathematically:

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

$$\text{Precision} = TP / (TP + FP)$$

$$\text{Recall} = TP / (TP + FN)$$

$$F1 = 2PR / (P + R)$$

Where TP = True Positive, TN = True Negative, FP = False Positive, FN = False Negative, P = Precision and R = Recall.

The models performance is shown in table 1.

Table 1. Models performance

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.8561	0.56	0.04	0.08
Random Forest	0.8443	0.24	0.03	0.06
K-Nearest Neighbors	0.8314	0.24	0.07	0.11

Artificial Intelligence, Machine Learning, and Data Science Journal

Research Article

Support Vector Machine	0.8538	0.00	0.00	0.00
CatBoost	0.8479	0.33	0.05	0.09
XGBoost	0.8231	0.21	0.08	0.12

Given these results, Logistic Regression emerges the most accurate at 85.61%, making it a better overall choice.

```

Logistic Regression Results:
Accuracy: 0.8561320754716981
Classification Report:
              precision    recall  f1-score   support

     0       0.86       0.99       0.92       725
     1       0.56       0.04       0.08       123

 accuracy          0.86          0.86          0.86          848
 macro avg       0.71       0.52       0.50          848
 weighted avg    0.82       0.86       0.80          848

Confusion Matrix:
[[721  4]
 [118  5]]
  
```

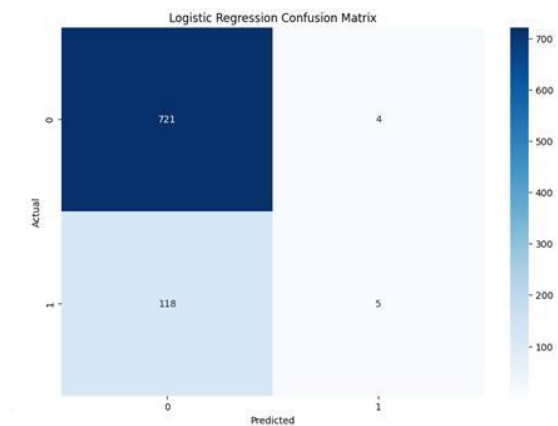


Figure 5: Logistics Regression Results

```

=====
Random Forest Results:
Accuracy: 0.8443396226415094
Classification Report:
              precision    recall  f1-score   support

     0       0.86       0.98       0.92       725
     1       0.24       0.03       0.06       123

 accuracy          0.84          0.84          0.84          848
 macro avg       0.55       0.51       0.49          848
 weighted avg    0.77       0.84       0.79          848

Confusion Matrix:
[[712 13]
 [119  4]]
  
```

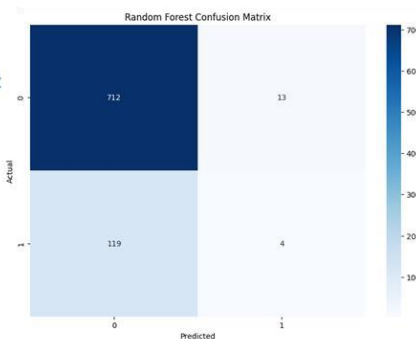


Figure 6: Random Forest Results

Artificial Intelligence, Machine Learning, and Data Science Journal

Research Article

K-Nearest Neighbors Results:
Accuracy: 0.8313679245283019
Classification Report:

	precision	recall	f1-score	support
0	0.86	0.96	0.91	725
1	0.24	0.07	0.11	123
accuracy			0.83	848
macro avg	0.55	0.52	0.51	848
weighted avg	0.77	0.83	0.79	848

Confusion Matrix:
[[696 29]
[114 9]]

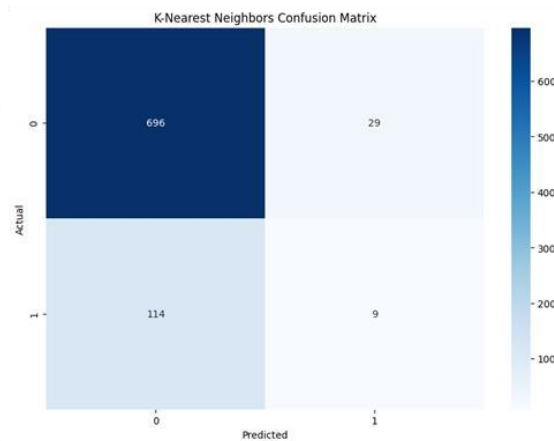


Figure 7: K-Nearest Neighbors Results

Support Vector Machine Results:
Accuracy: 0.8537735849056604
Classification Report:

	precision	recall	f1-score	support
0	0.85	1.00	0.92	725
1	0.00	0.00	0.00	123
accuracy			0.85	848
macro avg	0.43	0.50	0.46	848
weighted avg	0.73	0.85	0.79	848

Confusion Matrix:
[[724 1]
[123 0]]

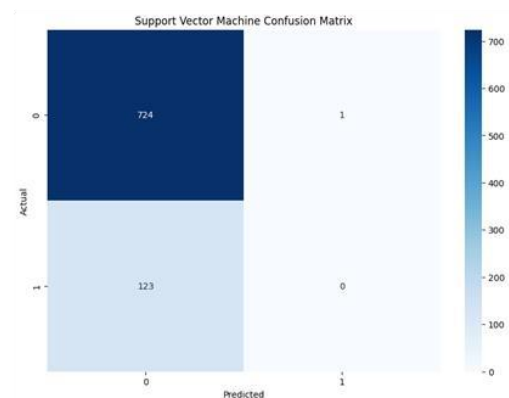


Figure 8: Support Vector Machine Results

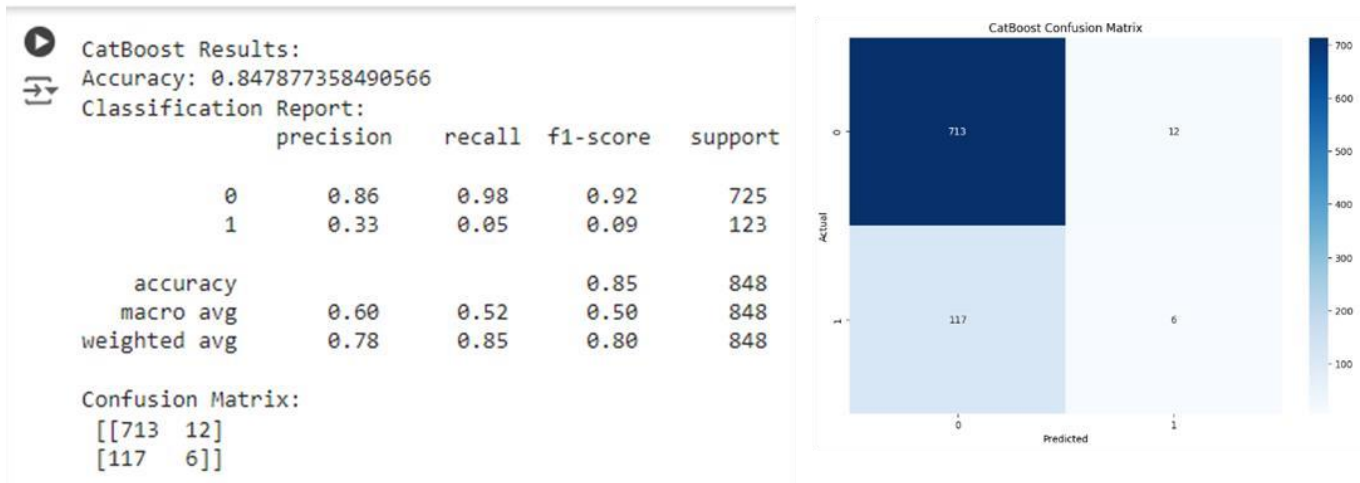


Figure 9: CatBoost Results

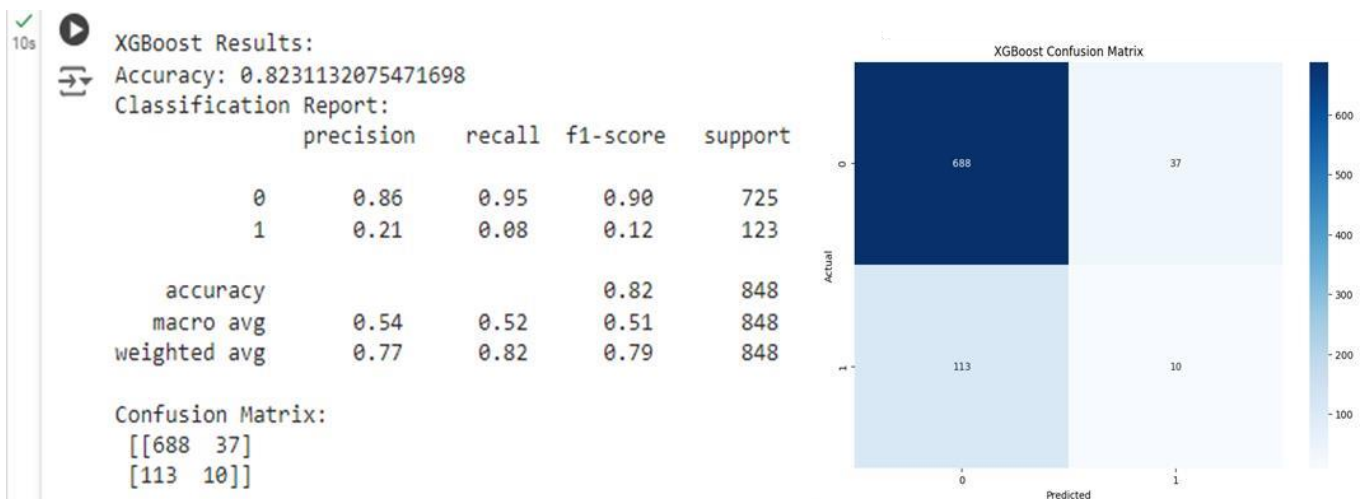
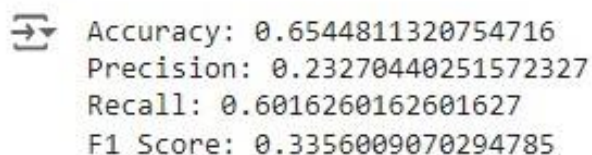


Figure 9: XGBoost Results

2.9 Addressing Class Imbalance With Smote

One of the most efficient techniques in overcoming class imbalance in a dataset is the Synthetic Minority Over-Sampling Technique, or SMOTE. This technique works by creating samples of the minority class to balance the distribution of classes. Class imbalance occurs when there are generally fewer instances in a single class compared

to other classes. This will make the model bias towards the majority class, usually perform badly in the prediction of the class that is less represented. Here, the positive class (presence of heart disease) represents the minority class, and it can be noticed that these models have a low recall. We apply SMOTE and reevaluate based on the logistic regression model to fix issues of class imbalance. This must help improve recall for the minority heart disease class.



```

➡ Accuracy: 0.6544811320754716
   Precision: 0.23270440251572327
   Recall: 0.6016260162601627
   F1 Score: 0.3356009070294785
  
```

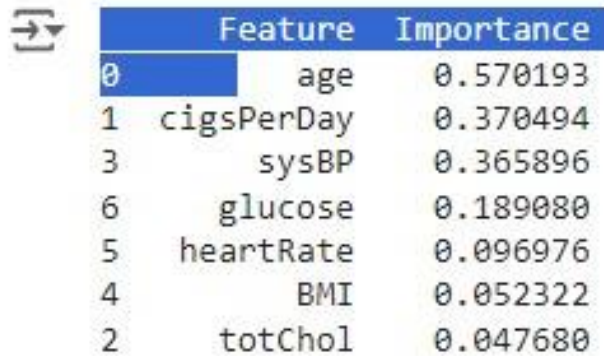
Figure 10: SMOTE Results

Interpretation of Results:

First, the model did improve significantly on the recall side, from 0.04 to 0.6016. Now, the model correctly identifies about 60% of heart diseases. The precision dropped to 0.2327 since most positive predictions are actually false positives. Reflecting this improvement is a balanced F1 score at 0.3356, even while accuracy decreased a bit to 0.6545. Although the accuracy has dropped, it is worth mentioning that accuracy alone may be misleading about the performance of the model, especially for class-imbalanced data sets. Probably, this focus on improving recall and F1 score means that the effectiveness of the model in heart disease case detection has improved significantly; this often counts more compared to mere optimization for overall accuracy.

3.0 Feature Importance Analysis

The following importance feature values are obtained from the Logistic Regression model and show the most influential features for heart disease predictions



	Feature	Importance
0	age	0.570193
1	cigsPerDay	0.370494
3	sysBP	0.365896
6	glucose	0.189080
5	heartRate	0.096976
4	BMI	0.052322
2	totChol	0.047680

Figure 11: Feature important Results

The analysis reveals that age (0.570193), daily cigarette consumption (0.370494), and systolic blood pressure (0.365896), among other risk factors, are very important in the prediction of heart disease, meaning that old, more cigarette-consuming, and higher systolic blood pressure persons have considerably higher risks. Glucose levels and heart rate showed a critical role of 0.189080 and 0.096976, respectively, stressing that controlling blood sugar levels and keeping heartbeats per minute at a healthy level is important for cardiac health. BMI and total cholesterol had minimum possible risk levels: 0.052322 and 0.047680 respectively; however, they contribute to the overall risk, meaning that both obesity and high levels of cholesterol have to be accounted for while designing heart disease prevention and control strategies.

3.0 RESULTS AND DISCUSSION:

From the analysis, Logistic Regression demonstrated the highest accuracy of 85.61% among the evaluated algorithms. It also achieved a balanced improvement in recall from 0.04 to 0.6016 and balancing the F1 score from 0.08 to 0.3356, which is crucial for accurately identifying heart disease cases while minimizing false positives. Critical features contributing to heart disease risk include age (0.570193), daily cigarette consumption (0.370494), and systolic blood pressure, these underscore the importance of lifestyle management and regular health monitoring in reducing cardiovascular risks. Additionally, glucose levels (0.189080), heart rate (0.096976), BMI (0.052322), and total cholesterol (0.047680) also play significant roles in predicting overall risk, emphasizing the need for maintaining optimal blood sugar levels and heart health to mitigate heart disease risks.

4.0 CONCLUSION

This research demonstrates that machine learning has colossal potential to ensure a change in cardiovascular health outcomes in Akwa Ibom State. Predictive models that implant critical indicators of health have been shown

to radically improve the rate of early detection, allowing interventions to be tailored to high-risk individuals. Not only do such tools raise the accuracy of diagnosis, but they equally equip diagnostic practitioners with proactive tools for preventive care.

5.0 RECOMMENDATIONS

The following recommendations are put forward based on findings of this study for healthcare practitioners, policymakers, and community stakeholders:

1. Embed machine learning-based diagnostics into routine healthcare process to improve the early detection capabilities in heart disease by sustained monitoring and risk stratification, which is personalized based on identified features of importance.
2. Design health interventions targeted at managing modifiable risk factors like smoking, raised blood pressure, or blood glucose. Institute smoking cessation programs and promote healthy lifestyle choices via outreach and education in the community.
3. Build an enabling environment by advocating policies that encourage the use of technology-driven healthcare solutions in Akwa Ibom State. This shall include incentivizing healthcare providers to embrace and integrate machine learning models into clinical decision-making.
4. Ensure that there are awareness programs about cardiovascular health management and health literacy at the community level. The project shall sensitize communities on the effects of lifestyle living on heart health and encourage constant checkups.
5. Invest in further research to refine machine learning algorithms specific to the local population in Akwa Ibom State. Collaborate with academic institutions and healthcare professionals to continuously improve predictive models and diagnostic accuracy.

REFERENCES

- Chawla, N.V., Bowyer, K.W., Hall, L.O., & Kegelmeyer, W.P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16, 321-357.
- Ekle, F., Bakpo, F., Udanor, C., & Eneh, A. (2023). A machine learning model and application for heart disease prediction using prevalent risk factors in Nigeria. *International Journal of Mathematical Analysis and Modelling*, 6(2).
- He, H., & Garcia, E.A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263-1284.

- Kebede Kassaw, A., Yimer, A., Abey, W. et al. (2023). The application of machine learning approaches to determine the predictors of anemia among under-five children in Ethiopia. *Scientific Reports*, 13(22919).
- Krittanawong, C., Zhang, H., Wang, Z., Aydar, M., & Kitai, T. (2017). Artificial intelligence in precision cardiovascular medicine. *Journal of the American College of Cardiology*, 69(21), 2657-2664.
- Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J.T. (2018). Deep learning for healthcare: review, opportunities and challenges. *Briefings in Bioinformatics*, 19(6), 1236-1246.
- Samuel, O., Zewotir, T., & North, D. (2024). Application of machine learning methods for predicting under-five mortality: analysis of Nigerian demographic health survey 2018 dataset. *BMC Medical Informatics and Decision Making*, 24(86).
- World Health Organization. (2021). Cardiovascular diseases (CVDs). Retrieved from [https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-\(cvds\)](https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)).