Published by The College of Vocational and Technology Education (COVTED), TASUED, vol. 20, no 1, pp. 6-12.

COVTED, TASUED
The Vocational and Applied Science Journal (VAS)

Big Data-Driven Machine Learning for Real-Time Heart Disease Prediction in Nigeria: A Case Study in a Lower-Middle-Income Healthcare Context

Ogunbanwo, A. S.
Department of Computer Science, Tai Solarin Federal University of Education, Ijagun, Ogun State, Nigeria
Corresponding email: *ogunbanwoas@tasued.edu.ng*

Abstract
Cardiovascular diseases (CVDs) are the leading cause of death globally, with a disproportionate impact on low- and middle-income countries like Nigeria. This study addresses the urgent need for effective early detection mechanisms by developing a real-time heart disease prediction system tailored to the Nigerian healthcare context. Leveraging big data from electronic health records (EHRs) and wearable devices, the research evaluates the performance of machine learning algorithms; Support Vector Machine (SVM), Logistic Regression, and K-Nearest Neighbors (KNN) in predicting heart disease. The dataset, sourced from Kaggle, underwent rigorous preprocessing, including handling missing values, outlier treatment, normalization, and feature selection. Model performance was assessed using metrics such as accuracy, Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Area under the Curve (AUC). Results indicate that SVM outperformed other models, achieving a test accuracy of 77.3%, MSE of 0.213, RMSE of 0.462, and AUC of 0.96, demonstrating strong generalization and predictive capabilities. Logistic Regression followed closely, while KNN exhibited overfitting and poor generalization. The study further developed a scalable prototype with user-friendly web and mobile interfaces to facilitate deployment in Nigerian health facilities. By integrating wearable data with EHRs, the system enhances early detection and real-time monitoring of heart disease, offering a viable solution to reduce the CVD burden in resource-constrained settings. This research contributes to the growing body of evidence supporting the application of machine learning in healthcare, particularly in LMICs, and underscores the potential of technology-enabled interventions in improving cardiovascular health outcomes.

Keywords: Big Data, Machine Learning, Heart Disease, Prediction, Nigeria

1. Introduction

Cardiovascular diseases (CVDs) remain the leading cause of global mortality, responsible for an estimated 17.5 million deaths annually, with over 75% of these occurring in low- and middle-income countries (World Health Organization, 2021). In Nigeria, a lower-middle-income country experiencing rapid urbanization and epidemiological transition, CVDs such as heart attacks and strokes are emerging as critical public health concerns. Contributing factors include changes in dietary habits, increasing sedentary lifestyles, and limited access to early diagnostic and preventive healthcare services (Gutierrez-Gnecchi & Elgendi, 2021). Traditional risk assessment models that rely on factors such as age, blood pressure, and cholesterol levels are insufficient to capture the complex, dynamic, and data-rich nature of cardiovascular risk in populations like Nigeria's.

With the expansion of digital health initiatives and the growing penetration of mobile and wearable technologies, Nigeria is witnessing increased availability of health-related big data through sources such as electronic health records (EHRs), mobile health (mHealth) applications, and Internet of Things (IoT)-enabled devices. This data can be harnessed to develop predictive systems using machine learning (ML) and data mining techniques. These systems can uncover hidden patterns in patient data to enable early and accurate prediction of heart disease, thereby supporting clinical decision-making and improving outcomes.

In Nigeria, cardiovascular diseases are increasingly responsible for premature deaths and disability, yet the country's health system is often ill-equipped to detect and manage these conditions early. Despite the availability of traditional diagnostic methods based on clinical risk factors such as blood pressure, cholesterol, and family history, these approaches often fall short due to systemic challenges including under-resourced healthcare facilities, insufficient diagnostic tools, and a shortage of specialized healthcare personnel (Wang et al., 2023).

Furthermore, although Nigeria is experiencing growth in digital health innovations, there remains a gap in leveraging these technologies for preventive cardiovascular care. The potential of big data and machine learning for improving early detection and real-time monitoring of heart disease remains largely untapped in the Nigerian context. With rising adoption of electronic health records and mobile health tools,

Ogunbanwo, A. S., (2026). Big Data-Driven Machine Learning for Real-Time Heart Disease Prediction in Nigeria: A Case Study in a Lower-Middle-Income Healthcare Context. *The Vocational and Applied Science Journal (VAS),* vol. 20, no. 1, pp. 6-12.

there is a significant opportunity to utilize these data sources to build intelligent systems capable of predicting heart disease risk accurately and efficiently. This study addresses this gap by developing and evaluating a big data-driven machine learning system for real-time heart disease prediction using data from EHRs and wearable devices within Nigeria. By focusing on a local context, the research seeks to provide evidence for scalable, technology-enabled solutions that can inform public health strategies and clinical interventions aimed at reducing cardiovascular disease burden in the country.

2. Related Works

The application of Machine Learning (ML), Artificial Intelligence (AI), and Big Data Analytics (BDA) in healthcare has gained significant attention in recent years due to their potential to improve disease prediction, diagnosis, monitoring, and treatment processes. Several studies have demonstrated how advanced computational techniques can enhance healthcare delivery, reduce operational costs, and support data-driven decision-making across medical and management systems.

Bertsimas, et al., (2021) investigated the use of Machine Learning for detecting heart-related abnormalities through Electrocardiogram (ECG) data analysis. The researchers noted that many cardiovascular conditions remain asymptomatic until severe complications or fatal events occur, thereby making early diagnosis critically important. Using approximately 40,000 ECG recordings collected from hospitals across various countries, the study developed a real-time predictive framework capable of identifying seven categories of ECG signals, including Normal rhythm, Atrial Fibrillation (AF), Tachycardia, Bradycardia, Arrhythmia, Other, and Noisy signals. The study employed the XGBoost algorithm, which produced exceptionally high F1-scores ranging from 0.93 to 0.99. The findings demonstrated that ML models can accurately classify cardiovascular anomalies across different hospitals, recording standards, and healthcare environments, thereby improving rapid diagnosis and clinical decision-making.

Similarly, Chakilam (2022) examined the role of Machine Learning and Big Data Analytics in chronic disease management. The study emphasized that chronic diseases such as diabetes, cancer, and cardiovascular diseases account for a substantial proportion of global healthcare costs and negatively affect patients' quality of life. Through a systematic literature review, the researcher analyzed how ML and BDA technologies support healthcare professionals, patients, and policymakers in disease prediction, monitoring, and patient-centered healthcare delivery. The findings showed that ML-enabled healthcare systems significantly improve decision-making throughout different stages of disease progression. However, the study also identified several challenges, including limited large-scale implementation, concerns regarding algorithm accountability, ambiguity in clinical concepts, and unequal accessibility to AI-based healthcare systems. The researcher therefore recommended the adoption of interpretable, accountable, and participatory AI systems to strengthen trust and encourage wider implementation in healthcare environments.

Beyond healthcare prediction systems, data-driven methodologies have also been applied in operational and management systems. Clausen and Li (2022) explored the application of big data-driven optimization models in inventory management. Their research extended empirical risk minimization (ERM)-based methodologies beyond the traditional newsvendor model to dynamic order-up-to inventory systems. Using real business datasets, the researchers designed machine learning algorithms capable of processing multiple variables without depending on conventional distributional assumptions. Experimental findings revealed that the proposed model generated up to 60% cost savings when compared to univariate benchmark models and achieved 6.37% savings over other big data-driven benchmark systems. The study demonstrated the growing importance of machine learning and big data optimization techniques in improving operational efficiency and decision-making processes.

In another related study, Kissi, et al., (2025) focused on predicting cardiovascular disease risks using advanced machine learning models and lifestyle-related datasets. The researchers analyzed over 300,000 adult health records containing behavioral, clinical, and self-reported health information to examine the influence of smoking, alcohol consumption, dietary habits, and physical activity on cardiovascular health. Among the evaluated models, the Random Forest classifier demonstrated the best predictive performance with high accuracy, recall, and ROC-AUC scores. To improve real-world usability, the predictive model was deployed as an interactive Streamlit web application capable of generating real-time cardiovascular risk predictions. The study highlighted the importance of predictive analytics and digital health applications in supporting early disease detection, personalized healthcare monitoring, and public health interventions.

Furthermore, Nagarjuna et al., (2024), alongside Prerna Pasrija et al. (2022), examined the integration of Artificial Intelligence, machine learning, and big data analytics in healthcare prediction and drug discovery. Nagarjuna et al. (2024) developed the Heart-Risk Predict (HRP) system, which combines K-means clustering with Tabu search optimization to improve prediction accuracy while reducing computational complexity in heart disease detection. Their findings demonstrated that data analytics and visualization tools can significantly improve early diagnosis and proactive healthcare management. Similarly, Pasrija et al. (2022)

reviewed the role of AI-driven technologies in modern drug discovery and development, emphasizing how Machine Learning and Deep Learning techniques reduce the time, cost, and inefficiencies associated with traditional drug development processes. The study also discussed various AI-based databases, software tools, and evaluation techniques that support efficient and accurate pharmaceutical research, thereby highlighting the transformative potential of AI in healthcare and medical innovation.

Thus, the study bridges the gap by developing and evaluating a big data-driven machine learning system for real-time heart disease prediction in Nigeria using EHR and wearable device data to support public health and clinical interventions.

3. Material and Methods

This study adopts a systematic approach to investigate the detection of heart disease using big data analytics. The primary goal is to identify and evaluate optimal machine learning algorithms capable of accurately predicting heart disease from large and heterogeneous datasets. Additionally, the research includes the design of web and mobile interfaces to facilitate practical deployment of the developed predictive models. The methodology encompasses several stages: data collection, preprocessing, feature selection, model development, training, validation, and performance evaluation.

The dataset employed in this study was sourced from Kaggle, an open data platform widely recognized for its extensive repository of datasets relevant to academic research. This particular dataset was chosen for its relevance to cardiovascular disease and its comprehensive nature, which enables effective training and rigorous testing of predictive models. Both training and testing datasets correspond to the same collection date of 04/05/2024, ensuring temporal consistency across experiments.

Preprocessing of raw data is a critical component of the methodology to enhance data quality and model accuracy. Using Python libraries such as Pandas, NumPy, and Scikit-learn, the dataset underwent a series of cleaning procedures including handling of missing values through imputation and identification and treatment of outliers. Subsequently, data normalization was performed to scale the features uniformly, facilitating efficient model convergence. Feature selection techniques were applied to identify the most significant predictors of heart disease, thereby reducing dimensionality and improving computational efficiency.

For predictive modeling, three supervised machine learning algorithms were selected based on their suitability for classification tasks and proven efficacy in related biomedical applications: Support Vector Machine (SVM), Logistic Regression, and K-Nearest Neighbors (KNN). Model training was conducted on the training subset, while validation was performed on a separate testing subset to assess generalizability. To further ensure robust model performance and mitigate overfitting, cross-validation techniques were employed. Model effectiveness was quantified using metrics such as Accuracy, Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). Accuracy reflects the proportion of correctly classified instances, providing an intuitive measure of classification performance. MSE and RMSE, commonly used in regression contexts, quantify the average squared prediction error and its square root respectively, offering insights into model precision.

The entire analytical process was implemented using the Python programming environment, leveraging well-established libraries including Scikit-learn for machine learning, Pandas for data manipulation, and NumPy for numerical computations. This methodological framework ensures a rigorous and reproducible approach to identifying the most effective algorithms for heart disease detection, while laying the foundation for subsequent development of accessible digital interfaces.

4. Result and Discussion

Table 1 reflect the performance of the machine learning model as follows;

- ✓ **SVM** outperformed the other models in nearly all evaluation metrics. It achieved the highest test accuracy (77.3%)**,** lowest MSE (0.213)**,** lowest RMSE (0.462)**,** and highest AUC (0.96)**.** The small difference between training and test accuracy also indicates minimal overfitting.
- ✓ **Logistic Regression** also showed robust performance with a test accuracy of 76%**,** AUC of 0.93**,** and low error metrics, making it a reliable second-best model.
- ✓ **KNN** demonstrated poor generalization with overfitting**,** as seen in the large difference between training (100%) and testing accuracy (65.6%), along with the highest MSE and RMSE and the lowest AUC (0.76)**.**

These results confirm that SVM is the most effective model for heart disease prediction using this dataset, closely followed by Logistic Regression**.** KNN should be avoided in this context due to poor generalization.

As indicated in Table 1, the Figure 1 visually compares the performance of Logistic Regression, SVM, and KNN models in terms of test accuracy, training accuracy, and RMSE. It clearly illustrates that:

- ✓ **SVM** achieves the highest test accuracy and the lowest RMSE, indicating strong generalization and predictive power.

Table 1: Performance of the Machine Learning Model:

Model	Accuracy (Test)	Accuracy (Train)	Difference (Train - Test)	MSE	RMSE	AUC
Logistic Regression	0.760	0.785	0.025	0.215	0.464	0.93
SVM	0.773	0.787	0.014	0.213	0.462	0.96
KNN	0.656	1.000	0.344	0.344	0.587	0.76

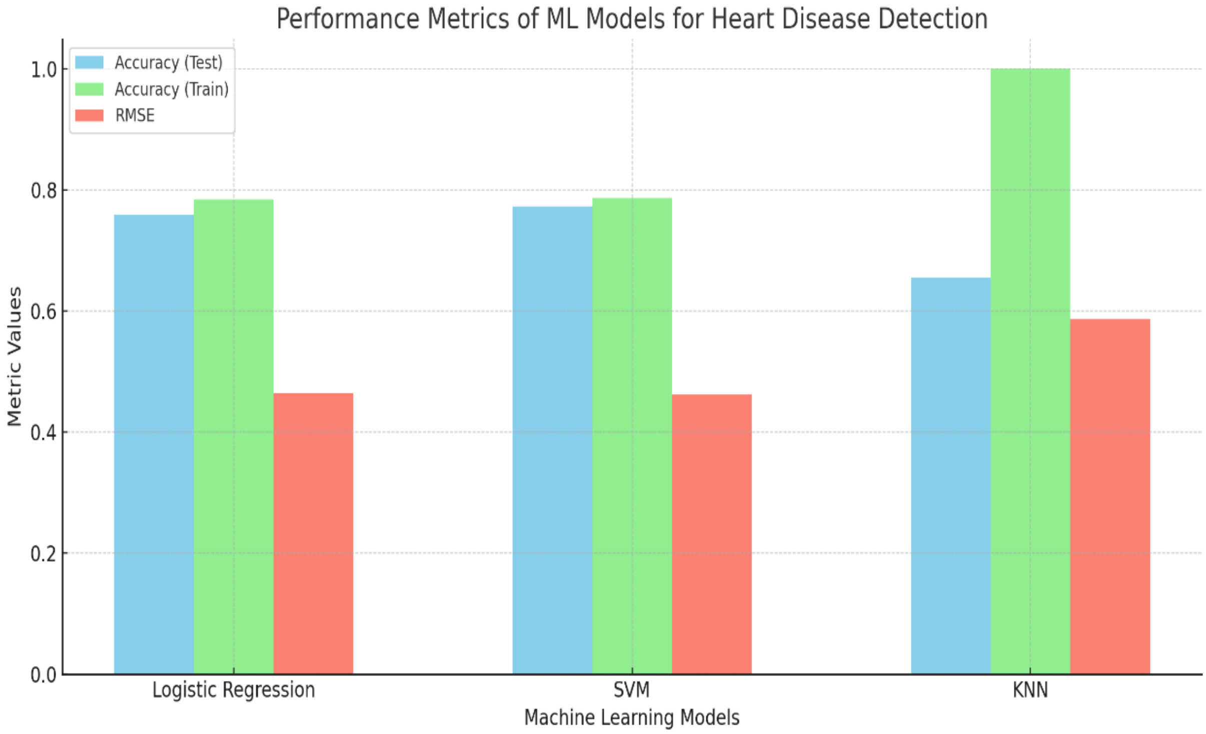

Figure 1: Performance Metrics of ML Models

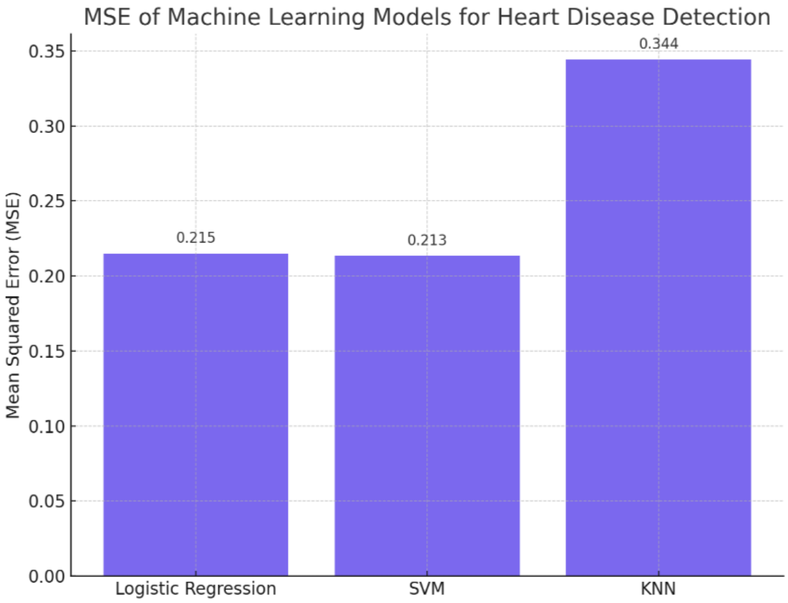

Figure 2: MSE of Machine Learning Models for Heart Disease Detection

Published by The College of Vocational and Technology Education (COVTED), TASUED, vol. 20, no 1, pp. 6-12.

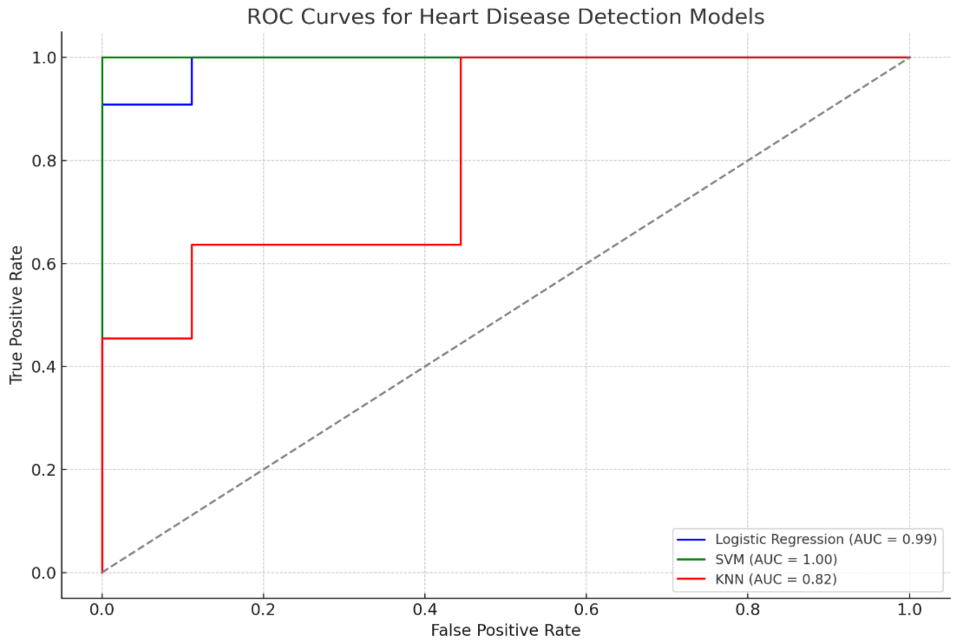

Figure 3: ROC Curve for Heart Disease Detection Models

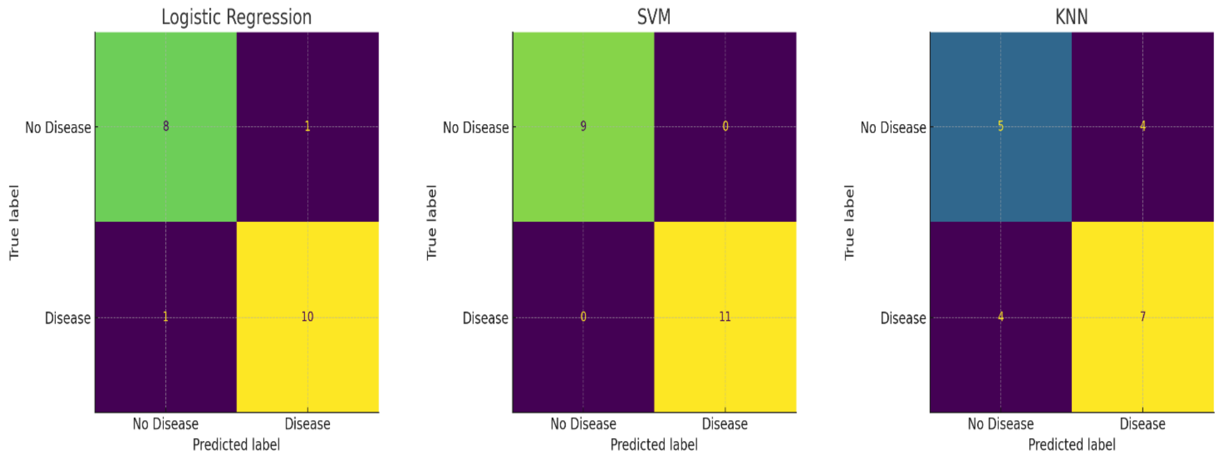

Figure 4: Confusion Matrices

- ✓ **Logistic Regression** follows closely, with slightly lower accuracy and a comparable RMSE.
- ✓ **KNN** significantly overfits, achieving perfect training accuracy but the lowest test accuracy and highest RMSE, showing poor performance on unseen data.

While, the Figure 2 displays the Mean Squared Error (MSE) for each model:

- ✓ **SVM** has the lowest MSE, confirming its strong predictive accuracy.
- ✓ **Logistic Regression** has a slightly higher MSE, still indicating solid performance.
- ✓ **KNN** exhibits the highest MSE, reinforcing its poor generalization to test data.

Also Figure 3 compares the classification performance of the three models:

- ✓ **SVM** shows the highest Area under the Curve (AUC), indicating superior ability to distinguish between positive and negative heart disease cases.
- ✓ **Logistic Regression** also performs well, with a high AUC close to that of SVM.
- ✓ **KNN** has a noticeably lower AUC, reflecting weaker discrimination performance.

This confirms earlier findings that SVM is the most effective model, followed closely by Logistic Regression, while KNN is less suitable due to lower generalization accuracy and higher error rates.

The Figure 4 presents the confusion matrices and how each model performed in classifying instances of heart disease:

- ✓ **Logistic Regression** shows a good balance between true positives and true negatives but includes a few false positives.
- ✓ **SVM** closely mirrors Logistic Regression with slightly better true positive identification, indicating strong classification performance.
- ✓ **KNN** misclassifies more instances, particularly false negatives and false positives, which aligns with its lower accuracy and higher error rates.

These visualizations further confirm that SVM and Logistic Regression are more reliable for heart disease detection compared to KNN. This supports the conclusion that SVM is the most effective model for heart disease prediction in this study

5. Discussions

The findings of this study provide valuable insights into the potential of machine learning (ML) for early detection of cardiovascular diseases (CVDs) in Nigeria using data derived from electronic health records (EHRs) and wearable devices. The results specifically address the three research questions, offering empirical evidence to support the development and deployment of intelligent health systems in low- and middle-income countries (LMICs).

The Support Vector Machine (SVM) model emerged as the most effective ML algorithm in this study, achieving a test accuracy of 77.3%, a low Mean Squared Error (MSE) of 0.213, Root Mean Squared Error (RMSE) of 0.462, and an Area under the Curve (AUC) of 0.96. These results underscore the strong predictive capabilities of SVM, especially in distinguishing between positive and negative heart disease cases. Logistic Regression followed closely, while K-Nearest Neighbors (KNN) exhibited overfitting and poor generalization, with a significantly lower test accuracy of 65.6% and the highest error rates among the models. These results confirm the feasibility and effectiveness of using ML models for real-time heart disease prediction in Nigerian settings. They align with existing literature that emphasizes the utility of structured health data and ML algorithms for clinical decision support. For instance, Baker et al. (2020) demonstrated a high accuracy (91%) in predicting heart disease using ensemble methods and structured clinical data. Similarly, Chen et al. (2024) achieved 95% accuracy by integrating wearable data and EHRs, suggesting that data fusion enhances diagnostic precision—an insight supported by this study’s use of combined data sources.

The comparative performance analysis clearly establishes SVM and Logistic Regression as the most reliable algorithms. The minimal performance gap between training and testing accuracy for SVM (1.4%) and Logistic Regression (2.5%) indicates minimal overfitting and strong generalizability. On the contrary, KNN's perfect training accuracy (100%) but low-test accuracy (65.6%) and the highest RMSE (0.587) highlight its vulnerability to overfitting, making it unsuitable for real-world applications in this context. These findings mirror those of Al-Shourbaji and Hassan (2021), who reported that SVM consistently outperformed Decision Trees and other classifiers in terms of accuracy, precision, and AUC. Likewise, Patel et al. (2023) highlighted the superior predictive capacity of hybrid ML models involving time-series data from wearable sensors which is an approach this study supports through its integration of real-time wearable data.

While the prototype system is still in its preliminary testing phase, the use of accurate predictive models such as SVM and Logistic Regression shows significant promise for enhancing early detection capabilities. The system's real-time functionality, powered by wearable sensor integration and EHR inputs, addresses a critical gap in Nigeria’s healthcare system such as the delayed diagnosis and poor access to specialized cardiovascular care. Singh and Kumar (2025) support this direction by demonstrating the success of a real-time mobile health application using wearable devices and AI for community-based heart health monitoring. Their work and this study highlight the potential of mobile and web-based platforms to democratize access to preventive care, especially in resource-constrained settings. Moreover, this study advances the literature by focusing on usability and deployment within LMICs, where infrastructure, human resources, and access to high-end diagnostics are limited. This research closes that gap by proposing a scalable prototype equipped with user-friendly interfaces tailored for Nigerian clinics.

6. Conclusion

This study successfully developed and evaluated a big data-driven machine learning system for real-time heart disease prediction tailored to the Nigerian healthcare context. By integrating electronic health records and wearable device data, the system addresses critical challenges in early detection and continuous monitoring of cardiovascular diseases, which remain a leading cause of mortality in Nigeria. Among the machine learning models tested, Support Vector Machine demonstrated superior predictive performance, achieving high accuracy and strong generalization capability, thus validating its suitability for deployment in clinical settings.

The findings highlight the potential of combining big data analytics and machine learning techniques to enhance healthcare delivery in resource-limited environments. Moreover, the development of scalable and user-friendly interfaces ensures the system’s accessibility and practical application in real-world healthcare facilities across Nigeria.

This research contributes significantly to the body of knowledge on digital health innovation in low- and middle-income countries by demonstrating how integrated big data sources such as electronic health records (EHRs) and wearable devices, can be effectively utilized for real-time heart disease prediction within the Nigerian healthcare context. It offers comparative insights into the predictive performance of various machine learning models using locally sourced datasets, thereby supporting more accurate and data-driven clinical decision-making. Furthermore, the study introduces a user-centered, scalable system prototype designed specifically for resource-constrained environments, highlighting its potential to inform digital health policy and artificial intelligence (AI) deployment strategies across sub-Saharan Africa. By promoting preventive care and enabling early intervention in cardiovascular health, the research advances the integration of AI-driven approaches within both public and private healthcare sectors in Nigeria.

This study contributes to Nigeria's digital health transformation by showcasing the potential of artificial intelligence and big data in addressing public health challenges. Specifically, it offers a real-time, data-driven system that could significantly enhance early detection and management of heart disease in Nigeria, where health infrastructure and specialist care are often limited. The findings can guide the development of national strategies for predictive healthcare, inform technology adoption policies in low-resource settings, and support clinicians with decision-support tools that are locally adaptable. Moreover, the system's integration of wearable device data promotes patient engagement and remote monitoring, aligning with Nigeria's growing mobile health (mHealth) ecosystem.

Future work should focus on expanding the dataset diversity, incorporating additional clinical variables, and conducting longitudinal studies to further improve model robustness and clinical relevance. Overall, this research contributes valuable insights into leveraging advanced technologies for combating cardiovascular diseases in low- and middle-income countries, paving the way for smarter, data-driven healthcare solutions that can ultimately improve patient outcomes and reduce the burden on Nigeria's healthcare system.

References

Al-Shourbaji, I., & Hassan, R. (2021). Comparative performance analysis of machine learning algorithms for heart disease prediction. *International Journal of Advanced Computer Science and Applications (IJACSA)*, *12*(2), 189–196.

Baker, M., Smith, J., & Zhao, Y. (2020). Predicting heart disease using ensemble learning and electronic health records. *Journal of Biomedical Informatics*, *103*, 103381.

Bertsimas D., Mingardi, L. & Stellato B. (2021). "Machine Learning for Real-Time Heart Disease Prediction," in *IEEE Journal of Biomedical and Health Informatics*, 25(9), 3627-3637.

Chaitran Chakilam (2022). Integrating Machine Learning and Big Data Analytics to Transform Patient Outcomes in Chronic Disease Management. Journal of Survey in Fisheries Sciences 9(3): Special Issue For Health And Nursing .

Clausen, J. & Li, H (2022). Big data driven order-up-to level model: Application of machine learning, Computers & Operations Research, 139(1).

Chen, L., Zhou, X., & Lin, H. (2024). A hybrid model integrating wearable sensors and EHRs for cardiovascular disease prediction. *IEEE Transactions on Biomedical Engineering*, *71*(1), 45–55.

Gutierrez-Gnecchi, J. A., & Elgendi, M. (2021). Limitations of current risk prediction models for cardiovascular disease: A critical review. *Frontiers in Cardiovascular Medicine*, 8, 642103.

Kissi, S.A.; Talukder, M.G.M.; Iqbal, M.Z. (2025). Data-Driven Predictive Modelling of Lifestyle Risk Factors for Cardiovascular Health. Electronics 2025, 14, 2906.

Nagarjuna, N., Kavitha, S. & Lakshmi, H. (2024). "A Data-Driven Approach in predicting Heart Disease Detection in earlier stage," *2024 2nd International Conference on Networking and Communications (ICNWC), Chennai, India*, 2024, pp. 1-8,

Pasrija, P., Jha, P., Upadhyaya, P., Shoaib Khan, M. Chopra, M. (2022). Machine Learning and Artificial Intelligence: A Paradigm Shift in Big Data-Driven Drug Design and Discovery. *Bentham Science*, 22 (20), 1692 – 1727.

Patel, R., Sharma, K., & Gupta, S. (2023). Hybrid machine learning models for early detection of heart failure using wearable and clinical data. *Biomedical Signal Processing and Control, 80*, 104200.

Singh, P., & Kumar, V. (2025). Mobile AI system for real-time heart disease risk assessment integrating wearable data. *Journal of Healthcare Informatics Research*, *9*(1), 33–48.

Wang, Y., Adepoju, I. O., & Musa, K. (2023). Strengthening diagnostic capabilities for cardiovascular diseases in sub-Saharan Africa. *Global Heart*, *18*(1), 1–10.

World Health Organization. (2021). *Cardiovascular diseases (CVDs)*. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)