Journal of Health Informatics in Africa

Enhancing Diabetes Mellitus Diagnosis and Management in Ghana through Supervised Learning: A Predictive Modeling Approach

Moses Peter Teye Ofoe[1], *Kate Takyi[1], Gadafi Iddrisu Balali[2], Boadu-Acheampong Samuelson[1], Rose-Mary Owusuaa Mensah Gyening[1]

[1]Department of Computer Science, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana.
[2]Department of Science, Seventh-day Adventist College of Education, Agona, Ghana.

Background: Diabetes mellitus is an emerging public health problem in Ghana where delayed diagnosis, limited specialized services and budget limitations may hinder patients from accessing prompt therapy. Machine learning based clinical decision support systems can help clinical judgement by recognizing patterns within routinely gathered patient data to provide timely diagnostic support based on evidence.
Purpose: In this work supervised machine learning models for the prediction of diabetes mellitus were created using clinical data from a healthcare institution in Ghana and the models were internally validated and their potential clinical decision support application discussed.
Methods: We acquired a retrospective data set of 3407 patient records from St Michael's Hospital, Jachie-Pramso, Ghana. The dataset contained eight predictor variables, including age, sex, fasting plasma glucose, glycated haemoglobin, nocturia, polyuria, weight loss and body mass index and ultimately a binary diabetes outcome. The data pre-processing includes removal of duplicate and missing entries, encoding of categorical variables and setting up the data set for modeling. The Random Forest, Gaussian Naive Bayes and Extreme Gradient Boosting models were constructed using 80:20 training-test split. The model performance was assessed by accuracy, class-specific precision, recall, F1-score and confusion matrices.
Results: The best accuracy (99.5%) was obtained by XGBoost, followed by Random Forest (98.9%) and Gaussian Naive Bayes (94.5%) on the held-out test dataset. XGBoost also presented the best blend of class-specific accuracy, recall and F1-score. These data revealed XGBoost as the best candidate model among the three algorithms studied. However, results are internal validation of a single healthcare center and need external confirmation before clinical adoption.
Conclusion: XGBoost provided the best internal prediction performance and may offer the computational basis for a diabetes-centric clinical decision support system in Ghana. However, before the model can be integrated into normal patient treatment, it has to be externally verified in geographically and clinically varied groups and prospectively examined for calibration, usability, fairness, and therapeutic utility.

Keywords: Diabetes Mellitus, Predictive Modeling, Supervised Learning, Healthcare Decision-making, Treatment Outcomes.

1 Introduction

Computer applications have become vital instruments to address human demands especially in public health [1]. Among these applications, Artificial Intelligence (AI) is one of the major forces pushing the growth and innovation in a variety of fields including healthcare [2]. Specifically, the incorporation of AI into Clinical Decision Support Systems (CDSS) has changed healthcare delivery, providing timely and tailored support to both physicians and patients [1]. Although affluent nations have made headway towards CDSS implementation this is not the case in resource restricted situations like Ghana [1]. In such scenario there is

*Corresponding author address: Department of Computer Science, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana Email: takyikate@knust.edu.gh
© 2026 JHIA. This is an Open Access article published online by JHIA and distributed under the terms of the Creative Commons Attribution Non-Commercial License. J Health Inform Afr. 2026;13(1):166-181. DOI: 10.12856/JHIA-2026-v13-i1-592

no doubt that there is an urgent need for intelligent decision-making tools to improve the delivery of healthcare. Diabetes mellitus is a chronic disease characterized by either deficient production of insulin or poor usage of insulin and is a major public health problem worldwide [3]. Diabetes is a growing burden with increasing incidence, particularly in sub-Saharan Africa, where access to specialist healthcare services is restricted [4].

The rising burden of diabetes in Ghana emphasizes the need for novel approaches to enhance its diagnosis and management [5, 6]. Although there is research on software-based treatments for diabetes management, there is a significant gap in the use of expert driven decision support systems for diagnosis [7, 8]. Furthermore, most of the research efforts have been directed to general prenatal care and not to the issues related to the diagnosis of diabetes [8]. This work seeks to fill these gaps by proposing, creating and deploying a Diagnostic Decision Support System (DSS) for the Ghanaian setting. The suggested system utilizes expert knowledge and clinical data to increase the accuracy, speed and efficiency of diabetes diagnosis, especially in the disadvantaged areas with limited access to specialized care. Importantly, this study goes beyond prior treatments by highlighting the incorporation of Clinical Practice Guidelines (CPGs) and expert opinions into the design of the CDSS, assuring its relevance and efficacy in the Ghanaian healthcare setting.

1.1 Clinical Decision Support and Predictive Modelling for Diabetes

Clinical decision support systems combine patient-specific information with clinical expertise to help offer healthcare workers with evaluations, warnings, forecasts or suggestions that may assist clinical decision-making [9]. Traditional CDSSs are often rule-based and predetermined, while modern systems increasingly include machine-learning algorithms that can detect complicated and nonlinear associations in electronic health data [10]. Machine learning-supported CDSSs in diabetes care have been examined for disease-risk prediction, screening, diagnosis, therapy selection, glycaemic monitoring and prediction of complications [11]. The potential relevance of such applications is high in scenarios where expert services are restricted and routinely acquired clinical information might be utilized to help frontline healthcare personnel.

Supervised machine-learning algorithms develop correlations between labeled predictor factors and a known clinical result [12]. Random Forest employs several decision trees and aggregated voting to increase the stability of the prediction and to lessen the danger of overfitting of decision trees [13]. Gaussian Naive Bayes estimates class membership by applying Bayes' theorem and conditional independence assumptions, whereas XGBoost creates decision trees in a sequential manner, where later trees fix errors caused by previous trees [14]. Studies on machine learning methods for diabetes prediction have revealed that ensemble methods like Random Forest and gradient boosting models usually surpass single classifiers [15, 16]. However, its documented performance is highly dependent on the population studied, specification of outcomes, sample size, preprocessing steps, validation method and class imbalance [17].

Previous studies highlighted the promise of machine learning for clinical decision assistance in diabetes. Islam et al. built a CDSS to predict type 2 diabetes and shown that the combination of clinical features and machine learning classification might assist in the early identification of patients at risk [18]. Similarly, Tasin et al. tested multiple algorithms using a privately gathered diabetes dataset and found that the prediction performance was highly dependent on feature selection, data quality, and the demographic represented in the dataset [19]. While such research shows the technical promise of machine learning, many of them are based on datasets that are regularly reused or populations outside sub-Saharan Africa. Therefore, their results cannot be automatically generalized to Ghana because of variations in patient characteristics, clinical practices, illness distribution, healthcare access and data-recording methods that may impact model performance.

In Ghana, digital interventions have focused mostly on diabetes education, lifestyle change, self-management, and patient-health care provider communication [20, 21]. Other Ghanaian research have focused on decision assistance tools in areas such as prenatal care [22]. However, information on supervised machine learning models constructed utilizing routinely obtained Ghanaian clinical data particularly for diabetes diagnosis is quite few. In addition, published predictive studies tend to focus just on accuracy and do not sufficiently evaluate class-specific sensitivity, precision, F1-score, model calibration, external validity or impact of false-negative predictions. This is significant because a model with a high overall accuracy might nevertheless miss clinically relevant cases, especially when the outcome classes are uneven.

© 2026 JHIA. This is an Open Access article published online by JHIA and distributed under the terms of the Creative Commons Attribution Non-Commercial License. J Health Inform Afr. 2026;13(1):166-181. DOI: 10.12856/JHIA-2026-v13-i1-592

The primary research gap that the current work addresses is the little knowledge on the comparative efficacy of supervised machine-learning algorithms trained on locally generated Ghanaian clinical data for diabetes prediction. In this study, the Gaussian Naive Bayes, Random Forest and XGBoost are compared using accuracy, precision, recall, F1-score and confusion matrices. It also determines the top performing candidate algorithm for additional validation and possible implementation into a Ghanaian diabetic CDSS. The study does not assess real clinical uptake or patient outcome but provides an internally validated prediction basis that needs external, prospective confirmation before deployment.

1.2 Theoretical and Conceptual framework

The Technology Acceptance Model gives a theory-based implementation perspective to understand the possible acceptance of the proposed CDSS [23]. The model indicates that users' adoption of a technology is mostly affected by two factors, perceived utility and perceived ease of use. Perceived utility in this context is defined as the degree to which healthcare professionals feel a predictive CDSS may enhance the timeliness, consistency and accuracy of diabetes-related clinical choices. Perceived ease of use was related to the extent to which doctors felt the system could be controlled and understood without undue effort or disruption to clinical workflow. These views may affect clinicians' attitudes, desire to utilize the system and ultimately routine use.

In the present study, the Technology Acceptance Model was not empirically tested, as the focus was on model creation and internal predictive performance, not on clinician conduct. The concept has consequences for future development of CDSS, nonetheless. However, a technically sound model may not provide therapeutic advantages if its predictions are hard to comprehend, inadequately embedded in the workflow or not accepted by healthcare personnel. Therefore, future implementation studies should measure perceived utility, convenience of use, trust, interpretability, fit with workflow and intention to use by potential healthcare consumers.

Conceptually, the investigation starts with regular demographic, biochemical, anthropometric and symptom-related information. These variables are preprocessed and three supervised learning algorithms are trained. The algorithms make binary predictions that are compared to the reported diabetes result using accuracy, precision, recall, F1-score and confusion matrices. The top performing candidate model may then be implemented in a CDSS which can offer risk estimations or diagnostic cues to health care practitioners (figure 1). The ultimate choice must always be made by clinical judgement and approved diagnostic protocols. The model is meant to help qualified healthcare professionals, not to substitute them.

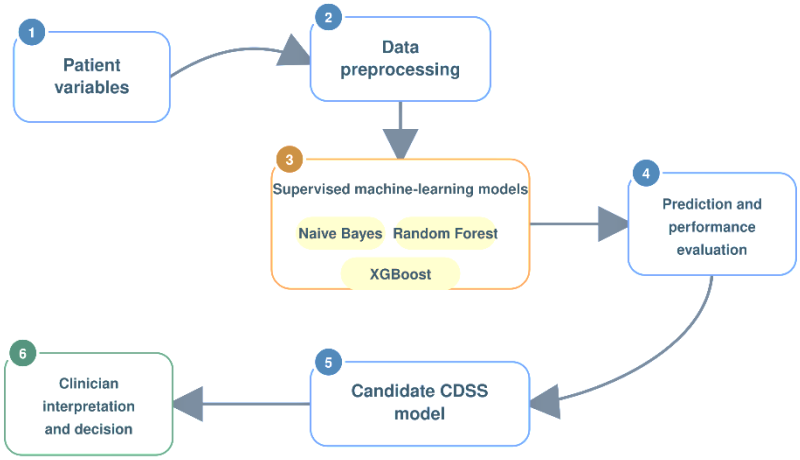

Fig. 1. Conceptual framework for supervised machine-learning-supported diabetes decision support.

The relevance of this study is more than academic questioning and practical use for healthcare professionals as well as patients. The proposed CDSS would provide general practitioners with enhanced decision-making tools and has the potential to fill the gap in diabetes treatment, particularly in rural locations with limited medical resources [24]. Ultimately this research helps to the continuing efforts to enhance healthcare delivery and decrease the socioeconomic burden of diabetes in Ghana and comparable

© 2026 JHIA. This is an Open Access article published online by JHIA and distributed under the terms of the Creative Commons Attribution Non-Commercial License. J Health Inform Afr. 2026;13(1):166-181. DOI: 10.12856/JHIA-2026-v13-i1-592

resource restricted contexts. The main purpose of this work was to create and internally validate supervised machine learning models for prediction of diabetes mellitus using clinical data collected from St. Michael's Hospital, Jachie-Pramso, Ghana. The specific aims of the present study were to (1) preprocess routinely collected clinical data for diabetes prediction, (2) build Gaussian Naive Bayes, Random Forest and XGBoost classification models, (3) compare the models using accuracy, precision, recall, F1-score and confusion matrices, and (4) identify the best-performing candidate model for further external validation and possible future integration into a diabetes-focused CDSS.

2 Methodology

2.1 Dataset Collection

In this study, the dataset used was collected from the St. Michael hospital from a community in the Ashanti region of Ghana called Jachie Pramso. The dataset contained 3407 instances with 2232 diabetic and the remaining 1175 being non-diabetic. The dataset has nine attributes with one as an output class with binary value indicating if a person is diabetic or not.

2.2 Data Preprocessing

The data was first preprocessed using an open-source data cleaning tool called openRefine to remove null values and redundant rows. It was further preprocessed by converting the True/False values to a binary 0 and 1 to allow it to be used for training by the model.

This is needed to train models like logistic regression, random forest, and xgb since they cannot use words for training. After all irrelevant and imbalanced features are dropped to prevent overfitting and increase the model's robustness.

2.3 Study Design

This research work is aimed at finding out the benefits of using clinical decision support to diagnose and manage diabetes in Ghana and getting the best model that can predict diabetes in people, Following the gathering and pre-processing of data, the dataset will be trained using supervised machine learning techniques to obtain the best model capable of predicting diabetes in people. The best model would be chosen and used to predict occurrence on the dataset. Figure 2 depicts how the desired outcome will be attained.

2.4 Data Description

The dataset used in this study was gathered from the St. Michael Hospital in Jachie Pramso, Ghana, in the Ashanti area. These data amounted to a total of three thousand four hundred and seven (3407) samples, divided into two groups: non-diabetic people and diabetes people. The dataset comprises physical examination indicators such as age, gender, fasting blood glucose level, glycated hemoglobin level, body mass index, and patient signs and symptoms (figure 3).

© 2026 JHIA. This is an Open Access article published online by JHIA and distributed under the terms of the Creative Commons Attribution Non-Commercial License. J Health Inform Afr. 2026;13(1):166-181. DOI: 10.12856/JHIA-2026-v13-i1-592

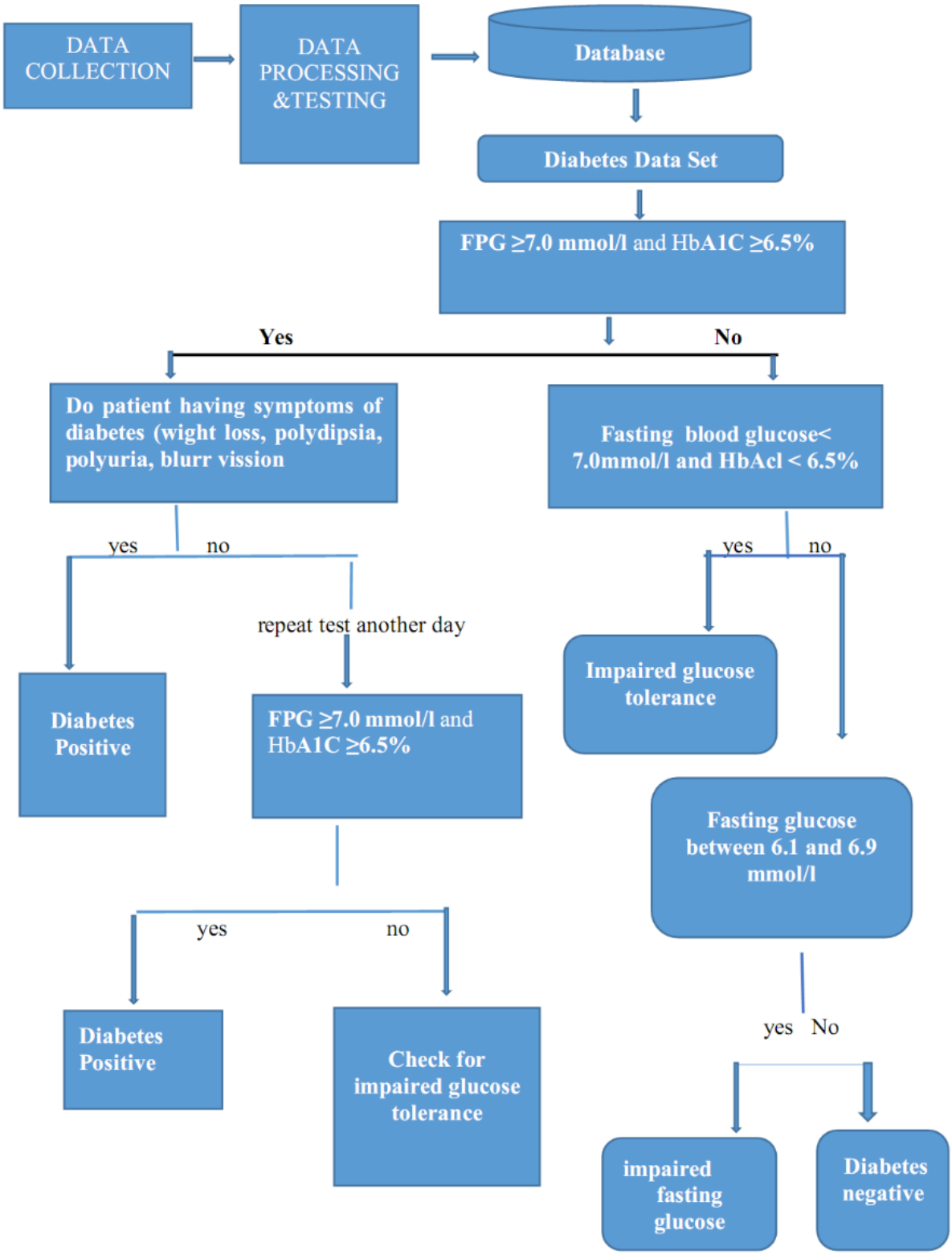

Fig. 2. Flow Chart for the Diagnostic Model

© 2026 JHIA. This is an Open Access article published online by JHIA and distributed under the terms of the Creative Commons Attribution Non-Commercial License. J Health Inform Afr. 2026;13(1):166-181. DOI: 10.12856/JHIA-2026-v13-i1-592

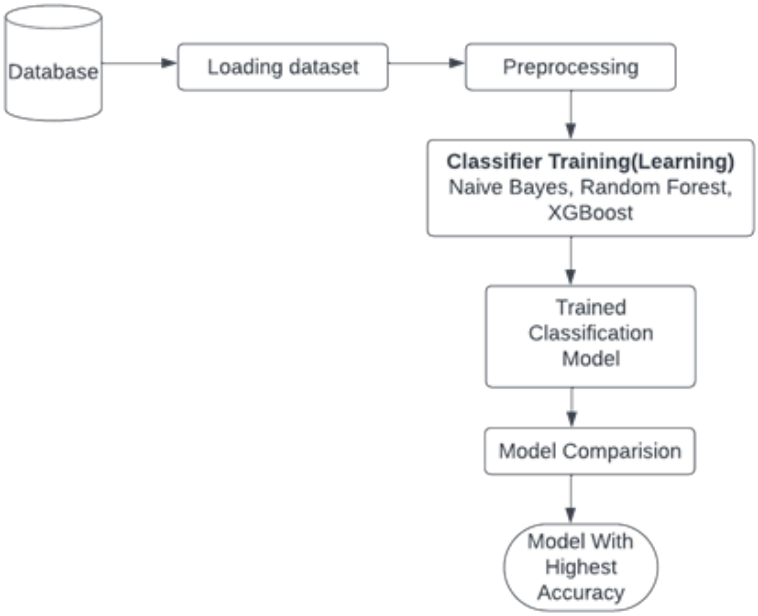

Fig. 3. Data Flow Diagram for the diagnostic model

2.5 Model Design Algorithms

2.5.1 The Naive Bayes algorithm

Naive Bayes is a probabilistic method based on Bayes theory. It is a Supervised Learning algorithm i.e. it trains on labeled data. Assumes features are independent. For example, the likelihood of a class label is independent of the values of all other attributes. This assumption is typically not true with real data but can be beneficial in many circumstances. It is an efficient and simple algorithm. It's capable of big data sets. It is also resistant to noise and outliers. This makes it ideally suited for medical data which is frequently inadequate and noisy. In our situation, we have employed the technique of Gaussian naive bayes which is more sensitive to data. The accuracy of the model for real-valued data was 94.57%.

2.5.2 Random Forest

The Random Forest is one of the most accurate and widely used ensemble machine learning methods. It operates by creating a huge number of decision trees throughout the training phase. Each tree is constructed using a random portion of the training data and a random subset of the features, which adds variety and helps prevent overfitting. During the prediction phase, the algorithm calculates the predictions of all the individual trees. It does this via majority voting in the case of classifying tasks and by averaging in the case of regressive tasks. The ensemble method enhances the prediction performance of the model, and generalizes to unknown data. Random Forest also gives you an idea of the relevance of features. It assesses how much each attribute contributes to the model's accuracy. This helps us comprehend the relationships in the data. The Random Forest is judged to be around 98.97% accurate. Random Forest finds use in banking, healthcare, natural language processing and many other domains.

2.5.3 XGBoost (Extreme Gradient Boosting)

Extreme gradient boosting (XGBoost) is a high-performance and high-efficiency machine learning technique for classification and regression, especially for structured data. XGB belongs to the gradient-boosting family and became famous because of its excellent predictive power and flexibility. XGBoost generates an ensemble of decision trees iteratively. The objective is to reduce mistakes in prediction by integrating predictions of already created trees to improve accuracy with each tree that is produced. This iterative process is built upon the optimization of gradient descent. XGB has built-in regularization to

© 2026 JHIA. This is an Open Access article published online by JHIA and distributed under the terms of the Creative Commons Attribution Non-Commercial License. J Health Inform Afr. 2026;13(1):166-181. DOI: 10.12856/JHIA-2026-v13-i1-592

prevent over-fitting. This provides XGBoost with robustness to noisy or difficult datasets. XGB has L1 and L2 regularization terms in objective function. The best accuracy of XGBoost is 99.56 %.

3 Results and Discussion

This work generated predictive models for diabetes mellitus categorization utilizing the Gaussian Naive Bayes, Random Forest and XGBoost algorithms. The dataset consisted of 3407 patient records and eight predictor variables: age, sex, fasting plasma glucose, glycated hemoglobin, nocturia, polyuria, weight loss and body mass index. The binary outcome variable was diabetes status and hence not included in the predictor factors. The dataset was preprocessed and then divided into training and test sets in an 80:20 ratio. Each method was trained on the identical training dataset and assessed individually on the held-out test dataset using accuracy, class-specific precision, recall, F1-score and confusion matrices. Figure 4 shows an example of the pre-processed data that was utilized to create the model.

	AGE	Gender	FPG	HbA1c	Noctouria	Polyuria	Weight_loss	BMI	Outcome
0	52	F	7.0	6.7	True	True	True	31.0	positive
1	57	M	10.8	8.0	False	False	False	32.0	positive
2	56	F	3.6	5.0	False	False	False	23.0	negative
3	32	F	7.5	6.8	False	False	False	31.0	positive
4	51	M	5.7	6.2	False	False	False	20.0	negative

Fig. 4. Sample of the pre-processed data using Python libraries in the Google Colab environment

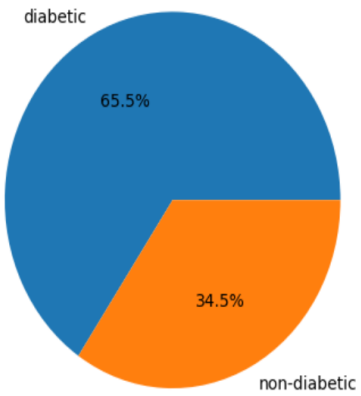

Fig. 5. Pie Chart showing a distribution of Diabetic and Non-Diabetic Patients

The pie chart visually represents the distribution of health conditions within the studied population (figure 5). It illustrates two main categories: "Diabetic" and "Non-Diabetic."

- Diabetic: This category occupies the majority of the chart, representing approximately 65.5% of the total population. It signifies individuals who have been diagnosed with diabetes.
- Non-diabetic: The remaining portion of the chart, accounting for approximately 34.5%, represents individuals who do not have diabetes.

This pie chart offers a clear and concise overview of the prevalence of diabetes within the research cohort, highlighting the significance of this health condition in the context of your study.

© 2026 JHIA. This is an Open Access article published online by JHIA and distributed under the terms of the Creative Commons Attribution Non-Commercial License. J Health Inform Afr. 2026;13(1):166-181. DOI: 10.12856/JHIA-2026-v13-i1-592

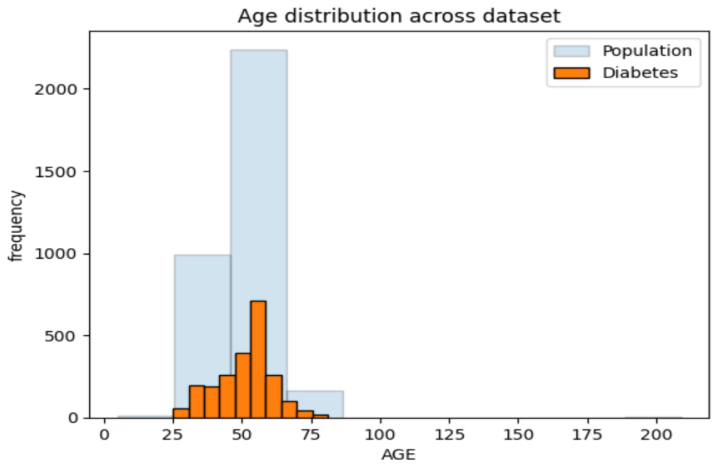

Fig. 6. Bar Chart showing Age distribution across Diabetic Patients

The figure 6 shows the age distribution of the people with diabetes across the dataset. The x-axis shows age in years, and the y-axis shows frequency, or the number of people in each age group.

The chart shows that the most common age group for people with diabetes is 65-74 years old. There is a gradual increase in the frequency of diabetes with age. The frequency of diabetes then declines slightly in the 75-84 age group and then more sharply in the 85+ age group.

Overall, the chart shows that diabetes is a more common condition in older adults. This is likely due to several factors, including age-related changes in metabolism, hormone levels, and immune function.

Summary of the age distribution of people with diabetes:

Most common age group: 65-74 years old
Slight decline in frequency in the 75-84 age group and a sharper decline in the 85+ age group
Diabetes is more common in older adults.

3.1 Results of Random Forest Algorithm

This model achieved an accuracy of 98.9%. The algorithm contained 100 trees improving the predictive accuracy and generalization of the model. The model's precision is 100% for non-diabetic patients and 99% for diabetic patients, the recall was 97% for non-diabetic and 100% for diabetic patients and an f1-score of 98% for non-diabetic and 99% for diabetic patients (figure 7).

```
Accuracy of model:  0.9897360703812317
              precision    recall  f1-score   support

           0       1.00      0.97      0.98       223
           1       0.99      1.00      0.99       459

    accuracy                           0.99       682
   macro avg       0.99      0.99      0.99       682
weighted avg       0.99      0.99      0.99       682
```

Fig. 7. Results from Random Forest Algorithm

The confusion matrix of the Random Forest algorithm shows that out of the 20% of data used as test data, 217/218 non-diabetic patients were correctly predicted with 1 non-diabetic predicted as diabetic. Out of 464 diabetic patients, 6 were predicted to be non-diabetic and the remaining 458 correctly predicted as diabetic (figure 8).

© 2026 JHIA. This is an Open Access article published online by JHIA and distributed under the terms of the Creative Commons Attribution Non-Commercial License. J Health Inform Afr. 2026;13(1):166-181. DOI: 10.12856/JHIA-2026-v13-i1-592

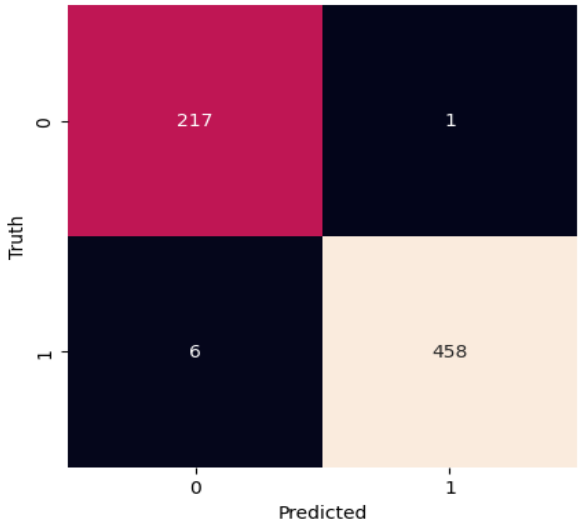

Fig. 8. Confusion Matrix for Random Forest Algorithm

3.2 Results of Naive Bayesian Algorithm

This model achieved an accuracy of 94.5%. The model's precision is 95% for non-diabetic patients and 94% for diabetic patients, the recall is 89% for non-diabetic and 98% for diabetic patients and an f1-score with 92% for non-diabetic and 96% for diabetic patients (figure 9).

Accuracy of model: 0.9457478005865103

	precision	recall	f1-score	support
0	0.95	0.89	0.92	235
1	0.94	0.98	0.96	447
accuracy			0.95	682
macro avg	0.95	0.93	0.94	682
weighted avg	0.95	0.95	0.95	682

Fig. 9. Results from Naive Bayesian Algorithm

The confusion matrix of the Gaussian Naive Bayes algorithm shows that out of the 20% of data used as test data, 208/218 non-diabetic patients were correctly predicted with 10 non-diabetic predicted as diabetic. Out of 464 diabetic patients, 27 were predicted to be non-diabetic and the remaining 437 correctly predicted as diabetic (figure 10).

© 2026 JHIA. This is an Open Access article published online by JHIA and distributed under the terms of the Creative Commons Attribution Non-Commercial License. J Health Inform Afr. 2026;13(1):166-181. DOI: 10.12856/JHIA-2026-v13-i1-592

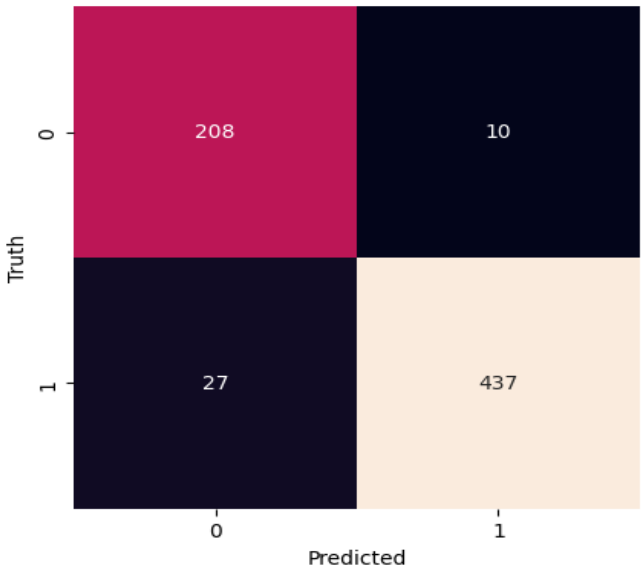

Fig. 10. Confusion Matrix for Naive Bayes Algorithm

3.3 Results of the XGBOOST Algorithm

This model achieved an accuracy of 99.5% by using 1000 estimators, a learning rate of 0.05% and 4 number of jobs. The model's precision is 100% for non-diabetic patients and 99% for diabetic patients, the recall is 98% for non-diabetic and 100% for diabetic patients and an f1-score with 99% for non-diabetic and 100% for diabetic patients (figure 11 and 12).

Accuracy of model: 0.9956011730205279

	precision	recall	f1-score	support
0	1.00	0.98	0.99	223
1	0.99	1.00	1.00	459
accuracy			0.99	682
macro avg	1.00	0.99	0.99	682
weighted avg	0.99	0.99	0.99	682

Fig. 11. Results from the XGboost Algorithm

© 2026 JHIA. This is an Open Access article published online by JHIA and distributed under the terms of the Creative Commons Attribution Non-Commercial License. J Health Inform Afr. 2026;13(1):166-181. DOI: 10.12856/JHIA-2026-v13-i1-592

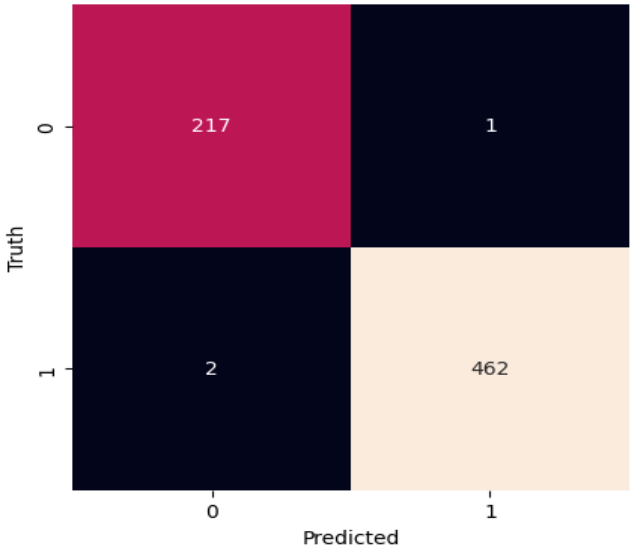

Fig. 12. Confusion Matrix for XGboost Algorithm

3.4 Comparison of Results for Algorithm Models

Table 1 examines the performance of 3 independently constructed candidate models on the identical held-out test dataset. The models were not ensembled together in single ensemble voting, stacking or averaging. Instead, Gaussian Naive Bayes and Random Forest were used as comparator models and the highest performing algorithm was chosen as the computational foundation of the proposed CDSS.

The total test accuracy of XGBoost was 99.5% which was more than Random Forest (98.9%) and Gaussian Naive Bayes (94.5%) . XGBoost also showed the greatest balanced performance per class. For non-diabetic class, it attained an accuracy of 100%, recall of 98% and F1-score of 99%. Precision, recall and F1-score for Diabetic class were 99%, 100% and 100% correspondingly. Random Forest showed similar performance with a little lower overall accuracy and non-diabetic recall. Gaussian Naive Bayes had the lowest accuracy and the lowest recall for non-diabetic patients, meaning it was more likely to misclassify non-diabetic patients.

Given these comparable results, XGBoost was selected as the final candidate prediction model for future validation and possible inclusion into a diabetes-oriented CDSS. In the proposed application, the model would take the eight patient characteristics and produce a binary prediction of diabetes or a risk indication linked with it for physician assessment. Random Forest and Gaussian Naive Bayes were kept as benchmark algorithms and no more predictions were added to the chosen XGBoost model. Hence, the “final model” refers only to the XGBoost candidate, not to a combination of the three techniques.

Table 1. Comparison of Results for Algorithm Models

Model	***Class of diabetes***	***Accuracy***	***Precision***	***Recall***	***F1 Score***
Random Forest	Non-Diabetic	98.9%	100%	97%	98%
	Diabetic		99%	100%	99%
XGBoost	Non-Diabetic	99.5%	100%	98%	99%
	Diabetic		99%	100%	100%
Naive Bayes	Non-Diabetic	94.5%	95%	89%	92%
	Diabetic		94%	98%	96%

Therefore, the link between the individual algorithm outcomes and the proposed CDSS is sequential, not additive. First, we trained and evaluated the three algorithms separately. Second, we analyzed their performance indicators to identify which algorithm most consistently differentiated between diabetes and

© 2026 JHIA. This is an Open Access article published online by JHIA and distributed under the terms of the Creative Commons Attribution Non-Commercial License. J Health Inform Afr. 2026;13(1):166-181. DOI: 10.12856/JHIA-2026-v13-i1-592

non-diabetic individuals. Third, XGBoost was picked as the candidate prediction engine for the greatest accuracy and the best trade-off between class-specific precision, recall and F1-score. The selected model, if externally verified, may be implemented in a CDSS interface that takes as input routinely gathered patient data, generates a prediction about diabetes and communicates the outcome to a trained physician. The output would inform screening and diagnostic prioritizing and judgments on confirmatory evaluation; it would not determine a diagnosis in its own right or replace authorized diagnostic criteria and professional judgement.

3.5 Comparison with prior diabetes prediction studies

The better performance of XGBoost in the current study is consistent with prior findings that tree-based ensemble and boosting algorithms tend to perform well on structured clinical data. XGBoost, unlike Gaussian Naive Bayes, does not assume conditional independence of predictors and may model non-linear connections and interactions between biochemical, anthropometric and symptom-related variables. Random Forest can also represent complicated interactions with the use of many decision trees, which may account for its near-accuracy to XGBoost and far higher accuracy than Gaussian Naive Bayes.

The XGBoost accuracy of 99.5% obtained in the present study was greater than the accuracy (81%), F1-score (0.81), and area under the curve (0.84) reported by Tasin et al. [19] for an XGBoost model created using a private diabetes dataset using an adaptive synthetic-sampling strategy. Supervised machine learning techniques can help type 2 diabetes prediction in a CDSS framework, however performance varies depending on the methodology, feature set and validation approach, as shown by Islam et al. These studies suggest the promise of supervised learning for diabetes-related decision assistance but also highlight that performance estimates are very dataset- and analysis design-dependent.

In a comprehensive review, Fregoso-Aparicio et al. [17] reported a high level of variation in diabetes-prediction studies with respect to demographics, predictor factors, algorithms and validation techniques. The research also found poor interpretability and inconsistent reporting as impediments to therapeutic applicability. Hence, the increased accuracy seen in the present investigation should not be construed as conclusive indication that the model would outperform previously reported models in other populations. The current dataset was collected from a single health institution in Ghana and consisted of very relevant indicators directly related to the clinical diagnosis of diabetes, i.e., fasting plasma glucose and HbA1c. Their inclusion may partly account for the very high performance.

Differences in diabetes prevalence, demographic makeup, sample size, missing data handling, feature selection, class imbalance and reference outcome definition between the current and prior research may potentially contribute to the observed differences. Moreover, the current estimations were based on a single 80:20 hold-out split, while several earlier research employed cross-validation or independent external validation. The results therefore give an optimistic internal evidence for the selection of XGBoost as a candidate model but require external and prospective validation before any inferences regarding generalisability or clinical superiority can be drawn.

3.6 Advantages of using clinical decision support to diagnose and manage diabetic patients

The performance measures reflect many attributes that would impact the potential utility and safety of the selected model in a CDSS. Accuracy measures the total proportion of properly categorized test observations, although accuracy alone is not sufficient for clinical assessment, especially given 65.5% of the research records were from patients classified as diabetes. Therefore, class-specific accuracy, recall and F1-score are more therapeutically meaningful data. The rounded reported result of 100% diabetes-class recalls shows that the XGBoost model was able to identify all, or almost all, diabetic instances in the held-out test dataset. High diabetic-class recall is particularly critical in a screening or diagnostic-support system, because false negative findings might delay confirmatory testing, treatment, and risk-factor management. The 99% diabetes-class precision shows that almost 99% of the records expected to be diabetic were classified as diabetic in the test dataset, indicating a low number of false-positive alarms. It may decrease unwanted referrals and further testing. A diabetic-class F1-score of 100% indicates a good mix of accuracy and recall.

For the non-diabetic class, XGBoost reached 100% accuracy, 98% recall and 99% F1 score. These results show considerable discrimination between the two outcome groups in the internal test sample. The balanced

© 2026 JHIA. This is an Open Access article published online by JHIA and distributed under the terms of the Creative Commons Attribution Non-Commercial License. J Health Inform Afr. 2026;13(1):166-181. DOI: 10.12856/JHIA-2026-v13-i1-592

class-specific performance of XGBoost is the key empirical reason for picking it as the potential predictive component of the proposed CDSS above Random Forest and Gaussian Naive Bayes. If these findings are confirmed by external data, the model might be used by doctors to identify patients who may need confirmatory examination, to prioritise individuals for additional review and to promote more uniform interpretation of routinely provided clinical information. However, the stated metrics assess predictive performance under retrospective internal testing settings. They do not show that the system improves patient outcomes, lowers waiting time, decreases death or improves diabetes control. Prospective studies are needed to analyze the clinical impact, workflow, usability and safety to assess such advantages.

3.7 Proposed integration of the candidate model into a CDSS

The present study identified XGBoost as a possible predictive component of a future diabetes-focused CDSS; it did not examine a fully implemented clinical system. In a future deployment, the relevant demographic, biochemical, anthropometric and symptom-related characteristics would be entered or retrieved by the physician using an electronic interface. The verified XGBoost model would then produce a prediction or indication of risk associated to diabetes, with the same pre-processing processes that were employed during model building. The system might show the prediction with important patient factors and cues for confirmation examination.

The forecast should be used in conjunction with recognized diagnostic criteria, clinical practice standards, the patient's medical history, and professional judgment. Thus, the suggested CDSS would act as a helpful tool rather than an independent diagnostic system. Before clinical deployment, the model needs to undergo external validation in other Ghanaian health settings, calibration assessment, performance evaluation across demographic and clinical subgroups, explainability analysis, prospective workflow testing, and assessment of its impact on improving clinical decisions and patient outcomes.

3.8 Testing and validating of decision support model

All three techniques were assessed internally on a held-out test subset of 20% of the available data, with the remaining 80% used for model training. All three models employed the same eight predictor variables and definition of binary outcome. The performance was evaluated by accuracy, precision, recall, F1-score and confusion matrices. XGBoost had the best performance on these criteria and was thus chosen as the potential prediction model.

This approach is an internal hold-out validation, and shows that the model worked well on records that were not used for training the model. However, it does not provide external validity across various hospitals, geographies or African people. Before the model can be regarded reliable for normal clinical application, the performance of the model has to be evaluated using independent datasets from healthcare institutions with various patient profiles, illness prevalence, clinical procedures and data-recording methods. Calibration, workflow compatibility, clinician acceptability, patient safety and clinical value also need prospective study.

4 Limitations of Study

There are various drawbacks to this study. First, a retrospective dataset was collected from one health center in Ghana. Thus, the sample may be reflective of the characteristics of the patients, clinical practices, illness prevalence and data collection techniques of that facility, and may not be representative of other hospitals, areas or populations. This restricts the generalisability of the findings outside and opens the risk of selection bias.

Secondly, the study was conducted using routinely collected clinical information. Information bias may have arisen from errors in measurement, data entry, diagnostic coding or documentation. Variables that could affect diabetes risk or diagnosis were not accessible for inclusion. These include family history, nutrition, physical activity, medication usage, comorbidities and socioeconomic features. There was also an imbalance between the diabetic and non-diabetic classifications, with the diabetes records comprising about 65.5% of the sample. Overall accuracy was recorded, but the imbalance could have impacted class-specific data.

© 2026 JHIA. This is an Open Access article published online by JHIA and distributed under the terms of the Creative Commons Attribution Non-Commercial License. J Health Inform Afr. 2026;13(1):166-181. DOI: 10.12856/JHIA-2026-v13-i1-592

Third, we added fasting plasma glucose and HbA1c as predictors. Including these parameters may have helped greatly to the very high prediction accuracy because the measurements are crucial to the clinical diagnosis of diabetes. There is a possibility of incorporation bias or target leakage if the recorded outcome was obtained using these tests. Future research should provide the definition of the reference diabetic outcome and consider models with and without variables used to determine that outcome.

Fourth, an internal review was carried out on a single 80:20 train-test split. The performance estimates from one split could be dependent on how the data are allocated. More robust estimates of model performance might be obtained using repeated stratified cross-validation, bootstrapping and independent external validation. Moreover, discrimination was not reported in the study employing area under the receiver operating characteristic curve, probability calibration, calibration plots, decision-curve analysis or clinically defined decision thresholds.

Finally, the study assessed predicting algorithms, not a fully deployed CDSS. It did not analyze model explainability, subgroup fairness, clinician acceptability, workflow integration, usability, cost effectiveness, patient safety, or impacts on clinical decision-making and health outcomes. Thus, the data should be seen as supportive evidence for selecting a candidate model for future validation and development, rather than as proof for a clinically successful CDSS.

5 Conclusion

In this work we created and internally validated Gaussian Naive Bayes, Random Forest and XGBoost models for the classification of diabetes mellitus using clinical data from St. Michael's Hospital, Jachie-Pramso, Ghana. The best held-out test accuracy was attained by XGBoost (99.5%), followed by Random Forest (98.9%) and Gaussian Naive Bayes (94.5%). XGBoost had the best overall mix of class-specific accuracy, recall and F1-score and was thus selected as the candidate prediction model for further validation and potential future incorporation into a diabetes-focused CDSS. The results reveal good internal predictive performance, but little clinical efficacy or generalisability. Before deployment, the XGBoost model must undergo recurrent internal validation, independent external validation, calibration and explainability assessment, subgroup fairness review, and prospective testing within normal clinical procedures. If the studies confirm its safety and efficacy, the model might help doctors identify patients who need additional examination and promote more consistent use of routinely obtained clinical information. Its output is intended to supplement, not replace, established diagnostic criteria, confirmatory testing, and expert clinical judgment.

6 Recommendations

- Healthcare institutions in Ghana should consider integrating the developed predictive models, particularly those utilizing XGBoost and Random Forest algorithms, into their clinical workflows. This integration can facilitate more accurate and timely diagnosis of diabetes mellitus, leading to improved patient outcomes.
- To maximize the potential of predictive modelling in healthcare decision-making, ongoing training and capacity-building programs should be provided to healthcare professionals. This includes training on how to interpret and utilize the output of predictive models effectively in clinical practice.
- Continuous efforts should be made to expand the availability and quality of healthcare data, particularly in the context of diabetes diagnosis and management. This may involve collaborations with healthcare facilities, research institutions, and government agencies to collect comprehensive and standardized datasets for model development and validation.
- Further validation and external evaluation of the developed predictive models are necessary to ensure their generalizability and reliability across different healthcare settings and populations. Collaborative studies with other healthcare institutions in Ghana and similar regions can provide valuable insights into the performance of these models in diverse contexts.
- Building upon the success of predictive modelling, the development and implementation of a robust Decision Support System (DSS) tailored to diabetes diagnosis and management in Ghana

© 2026 JHIA. This is an Open Access article published online by JHIA and distributed under the terms of the Creative Commons Attribution Non-Commercial License. J Health Inform Afr. 2026;13(1):166-181. DOI: 10.12856/JHIA-2026-v13-i1-592

should be prioritized. This DSS should incorporate the validated predictive models as well as clinical guidelines to provide actionable insights to healthcare providers at the point of care.
- Community engagement and education initiatives should be established to raise awareness about the importance of early detection and management of diabetes mellitus. This includes educational campaigns targeting both healthcare professionals and the general public to promote preventive measures and lifestyle modifications.

Acknowledgments

We express our sincere gratitude to the hospital administration and staffs of St. Michael's Hospital, Jachie-Pramso, for their invaluable support, cooperation, and provision of resources throughout the duration of this study. We also extend our appreciation to all individuals who contributed in various capacities to the successful completion of this work.

Conflict of Interest Statement

The authors declare that there are no conflicts of interest, financial or otherwise, regarding the publication of this paper. All authors have read, approved, and agreed to the submission of the final manuscript.

Author Contributions

- **Moses Peter Teye Ofoe :** Conceptualization, methodology, data collection, conduct of the study, investigation, and formal analysis.
- **Kate Takyi:** Resources, supervision, and project administration.
- **Gadafi Iddrisu Balali:** Drafting, critical review, and substantive revision of the manuscript, including response to reviewer comments.
- **Boadu-Acheampong Samuelson:** Data curation, methodological support, and draft review.
- **Rose-Mary Owusuaa Mensah Gyening:** Writing (original draft preparation), software, visualization, and
- formal analysis.

References

[1] Job, V., Framework of a Clinical Decision Support System for diagnosis of Type 1 diabetes. 2019.
[2] Tandon, P., Clinical Decision Support System for the Diagnosis, Analysis and Management of Hepatitis C. 2020.
[3] Amendt, T., et al., Autoreactive antibodies control blood glucose by regulating insulin homeostasis. Proceedings of the National Academy of Sciences, 2022. **119**(6): p. e2115695119.
[4] WHO, Report of the second meeting of the WHO Technical Advisory Group on Diabetes: virtual meeting, 16-17 December 2021. 2022.
[5] Naous, E., et al., Quality of life and other patient-reported outcomes in adult Lebanese patients with type 2 diabetes during COVID-19 pandemic. Journal of Endocrinological Investigation, 2022. **45**(4): p. 763-772.
[6] Tagoe, E.T., et al., Challenges to the delivery of clinical diabetes services in Ghana created by the COVID-19 pandemic. Journal of Health Services Research & Policy, 2022: p. 13558196221111708.
[7] Sowah, R.A., et al., Design and development of diabetes management system using machine learning. International journal of telemedicine and applications, 2020. **2020**.
[8] Abejirinde, I.-O.O., et al., Viability of diagnostic decision support for antenatal care in rural settings: findings from the Bliss4Midwives Intervention in Northern Ghana. Journal of global health, 2019. **9**(1).
[9] Sutton, R.T., et al., An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ digital medicine, 2020. **3**(1): p. 17.
[10] Berge, G.T., et al., Combining unsupervised, supervised and rule-based learning: the case of detecting patient allergies in electronic health records. BMC Medical Informatics and Decision Making, 2023. **23**(1): p. 188.

© 2026 JHIA. This is an Open Access article published online by JHIA and distributed under the terms of the Creative Commons Attribution Non-Commercial License. J Health Inform Afr. 2026;13(1):166-181. DOI: 10.12856/JHIA-2026-v13-i1-592

[11] Tuppad, A. and S.D. Patil, Machine learning for diabetes clinical decision support: a review. Advances in Computational Intelligence, 2022. **2**(2): p. 22.
[12] Lo Vercio, L., et al., Supervised machine learning tools: a tutorial for clinicians. Journal of Neural Engineering, 2020. **17**(6): p. 062001.
[13] Salman, H.A., A. Kalakech, and A. Steiti, Random forest algorithm overview. Babylonian Journal of Machine Learning, 2024. **2024**: p. 69-79.
[14] Deka, P.P. and J. Weiner, XGBoost for Regression Predictive Modeling and Time Series Analysis: Learn how to build, evaluate, and deploy predictive models with expert guidance. 2024: Packt Publishing Ltd.
[15] Malik, A., Comparative Evaluation of Ensemble Learning Algorithms for Early Detection of Diabetes. EDRAAK, 2025. **2025**: p. 103-110.
[16] Saihood, Q. and E. Sonuc, A practical framework for early detection of diabetes using ensemble machine learning models. Turkish Journal of Electrical Engineering and Computer Sciences, 2023. **31**(4): p. 722-738.
[17] Fregoso-Aparicio, L., et al., Machine learning and deep learning predictive models for type 2 diabetes: a systematic review. Diabetology & metabolic syndrome, 2021. **13**(1): p. 148.
[18] Islam, R., et al., Clinical decision support system for diabetic patients by predicting type 2 diabetes using machine learning algorithms. Journal of Healthcare Engineering, 2023. **2023**(1): p. 6992441.
[19] Tasin, I., et al., Diabetes prediction using machine learning and explainable AI techniques. Healthcare technology letters, 2023. **10**(1-2): p. 1-10.
[20] Iwelomen, O., et al., Therapeutic patient education programs on diabetes in sub-Saharan Africa: A systematic review. Plos one, 2024. **19**(6): p. e0299526.
[21] Nam, M.A., Examining The Role Of Self-Care Mangement In Type 2 Diabetes Mellitus Patients At The Korle Bu Teaching Hospital. 2017.
[22] Abejirinde, I.-O.O., et al., Pregnant women's experiences with an integrated diagnostic and decision support device for antenatal care in Ghana. BMC pregnancy and childbirth, 2018. **18**(1): p. 209.
[23] Davis, F.D., Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS quarterly, 1989. **13**(3): p. 319-340.
[24] Asante, K.T., Residual capacity and the political economy of pandemic response in Ghana. 2022, World Institute for Development Economic Research (UNU-WIDER).

© 2026 JHIA. This is an Open Access article published online by JHIA and distributed under the terms of the Creative Commons Attribution Non-Commercial License. J Health Inform Afr. 2026;13(1):166-181. DOI: 10.12856/JHIA-2026-v13-i1-592