Original Paper

# Machine Learning–Based Hyperglycemia Prediction: Enhancing Risk Assessment in a Cohort of Undiagnosed Individuals

Kolapo Oyebola[1,2], DPhil; Funmilayo Ligali[1,2], MS; Afolabi Owoloye[1,2], MS; Blessing Erinwusi[2], MS; Yetunde Alo[2], MS; Adesola Z Musa[1], DPhil; Oluwagbemiga Aina[1], DPhil; Babatunde Salako[1], Dr med

[1]Nigerian Institute of Medical Research, Lagos, Nigeria

[2]Centre for Genomic Research in Biomedicine, Mountain Top University, Ibafo, Nigeria

**Corresponding Author:**

Kolapo Oyebola, DPhil
Nigerian Institute of Medical Research
6, Edmund Crescent, Yaba
Lagos, 101245
Nigeria
Phone: 234 8034778549
Email: oyebolakolapo@yahoo.com

**Related Articles:**
Preprint (medRxiv): https://www.medrxiv.org/content/10.1101/2023.11.22.23298939v1
Preprint (JMIR Preprints): http://preprints.jmir.org/preprint/56993
Peer-Review Report by Tarek Abd El-Hafeez (Reviewer K): https://med.jmirx.org/2024/1/e60393
Peer-Review Report by Fakhare Alam (Reviewer V): https://med.jmirx.org/2024/1/e60389
Peer-Review Report by Akhil Chaturvedi (Reviewer AD): https://med.jmirx.org/2024/1/e60853
Author's Response to Peer-Review Reports: https://med.jmirx.org/2024/1/e60174

## Abstract

**Background:** Noncommunicable diseases continue to pose a substantial health challenge globally, with hyperglycemia serving as a prominent indicator of diabetes.

**Objective:** This study employed machine learning algorithms to predict hyperglycemia in a cohort of individuals who were asymptomatic and unraveled crucial predictors contributing to early risk identification.

**Methods:** This dataset included an extensive array of clinical and demographic data obtained from 195 adults who were asymptomatic and residing in a suburban community in Nigeria. The study conducted a thorough comparison of multiple machine learning algorithms to ascertain the most effective model for predicting hyperglycemia. Moreover, we explored feature importance to pinpoint correlates of high blood glucose levels within the cohort.

**Results:** Elevated blood pressure and prehypertension were recorded in 8 (4.1%) and 18 (9.2%) of the 195 participants, respectively. A total of 41 (21%) participants presented with hypertension, of which 34 (83%) were female. However, sex adjustment showed that 34 of 118 (28.8%) female participants and 7 of 77 (9%) male participants had hypertension. Age-based analysis revealed an inverse relationship between normotension and age ($r$=–0.88; $P$=.02). Conversely, hypertension increased with age ($r$=0.53; $P$=.27), peaking between 50-59 years. Of the 195 participants, isolated systolic hypertension and isolated diastolic hypertension were recorded in 16 (8.2%) and 15 (7.7%) participants, respectively, with female participants recording a higher prevalence of isolated systolic hypertension (11/16, 69%) and male participants reporting a higher prevalence of isolated diastolic hypertension (11/15, 73%). Following class rebalancing, the random forest classifier gave the best performance (accuracy score 0.89; receiver operating characteristic–area under the curve score 0.89; $F_1$-score 0.89) of the 26 model classifiers. The feature selection model identified uric acid and age as important variables associated with hyperglycemia.

**Conclusions:** The random forest classifier identified significant clinical correlates associated with hyperglycemia, offering valuable insights for the early detection of diabetes and informing the design and deployment of therapeutic interventions. However, to achieve a more comprehensive understanding of each feature's contribution to blood glucose levels, modeling additional relevant clinical features in larger datasets could be beneficial.

*JMIRx Med 2024;5:e56993;* doi: 10.2196/56993

**Keywords:** hyperglycemia; diabetes; machine learning; hypertension; random forest

# Introduction

Noncommunicable diseases (NCDs) have become a substantial public health concern in Africa [1]. Conditions like coronary artery disease, stroke, hypertension, and diabetes, which were once primarily associated with high-income nations or affluence, have now become pervasive health challenges in low- and middle-income countries and across diverse socioeconomic strata [1]. The complex nature of NCDs underscores the need for a comprehensive approach to risk assessment, intervention, and prevention.

Suburban communities serve as a distinctive microcosm within an evolving landscape of diseases [2,3]. These communities, characterized by the coexistence of traditional and modern lifestyles, grapple with risk factors that necessitate thorough examination [4]. The epidemiological shift from communicable to NCDs, coupled with limited health care resources, especially in suburban parts of low- and middle-income countries [5,6], stresses the importance of this research. In addition, recent advancements in genetic research have elucidated the underlying mechanisms of various complex NCDs. The identification of individuals at an elevated genetic risk for NCDs has the potential to revolutionize the approach of health care stakeholders to disease management. However, the effective implementation of genetic screening for NCD risk analysis relies on a robust understanding of the baseline contributors prevalent in the target population [7,8]. This study provided a comprehensive description of the prevalence and intricate interplay of risk factors associated with NCDs, highlighting hypertension, obesity, and diabetes. The specific focus was on undiagnosed individuals who were asymptomatic to elucidate the complex relationships of these health indicators within this population.

Machine learning encompasses a diverse set of algorithms designed to extract patterns from data and establish associations between these patterns and discrete sample classes within the data. Machine learning proves to be a valuable tool for identifying potential disease risk factors, elucidating etiology, and interpreting complex pathological processes in the context of NCDs [9-16]. In this study, multiple machine learning algorithms were developed to predict elevated blood glucose levels in a cohort of undiagnosed individuals who were asymptomatic. The primary objective was to systematically compare the accuracies of supervised machine learning classifiers to identify the most effective model for predicting hyperglycemia. Leveraging the predictors in the dataset, we meticulously constructed and evaluated these models for the identification of significant features associated with potential diabetes in the population.

# Methods

## Ethical Considerations

Ethical approval was obtained from the institutional review board of the Nigerian Institute of Medical Research (IRB/21/074). Data collected from participants was anonymized, and personal identifiers were removed. Furthermore, participants' data were stored in our database with access restricted to authorized research personnel only. The study participants received refreshments as compensation for their time and contribution. This gesture was intended to acknowledge their involvement and ensure their comfort during the study sessions while maintaining fairness and transparency in the compensation process.

## Participant Recruitment and Screening

This study was carried out as part of a parallel community-based genetic screening of apparently healthy adults living in Ijede Community, Lagos, Nigeria. Following informed consent, participants were recruited, and 10 ml of venous blood samples were collected per participant. Demographic information, BMI, knowledge, attitude, and practices were obtained from the participants. The study clinician also obtained the participants' personal and family medical history as well as their smoking status. Exclusion criteria included pregnancy at the time of recruitment, placement on antihypertensive or antidiabetic chemotherapy or radiotherapy, current or previous hematologic or tumoral diseases, and known chronic diseases. Participants underwent electrocardiogram (ECG) screening (SonoHealth, United States) to provide clues on heart defects or other heart-related problems. Hemoglobin electrophoresis was conducted to detect possible hemoglobinopathy in the participants [17]. In addition, random blood glucose (RBG) concentrations (Guilin Royalze, China) and blood pressure (BP) values (Iston Mediq, United States) were determined to evaluate the presence or absence of prediabetes, diabetes, prehypertension, or hypertension onset in the participants. Participants with screening tests outside normal ranges were advised to visit their health care specialists for further checks. Normal BP was described as systolic BP (SBP) <120 mmHg and diastolic BP (DBP) <80 mmHg. Elevated BP was defined as SBP 120-129 mmHg and DBP <80 mmHg, stage 1 hypertension (prehypertension) as SBP ≥130-139 mmHg and DBP 80-89 mmHg, and stage 2 hypertension as SBP ≥140 and DBP ≥90 mmHg [18]. Isolated systolic hypertension (ISH) was described as SBP >140 mmHg and DBP <90 mmHg [19]. Isolated diastolic hypertension (IDH) is an important subtype of hypertension defined as SBP <130 mmHg and DBP ≥80 mmHg [20]. Prediabetes was defined as an RBG concentration of 140-199 mg/dl or fasting blood glucose of 100-125 mg/dl. Diabetes mellitus was defined as an RBG level ≥200 mg/dl or fasting blood glucose level ≥126 mg/dl [21]. However, as all the

participants reported that they were not fasting, RBG values were documented.

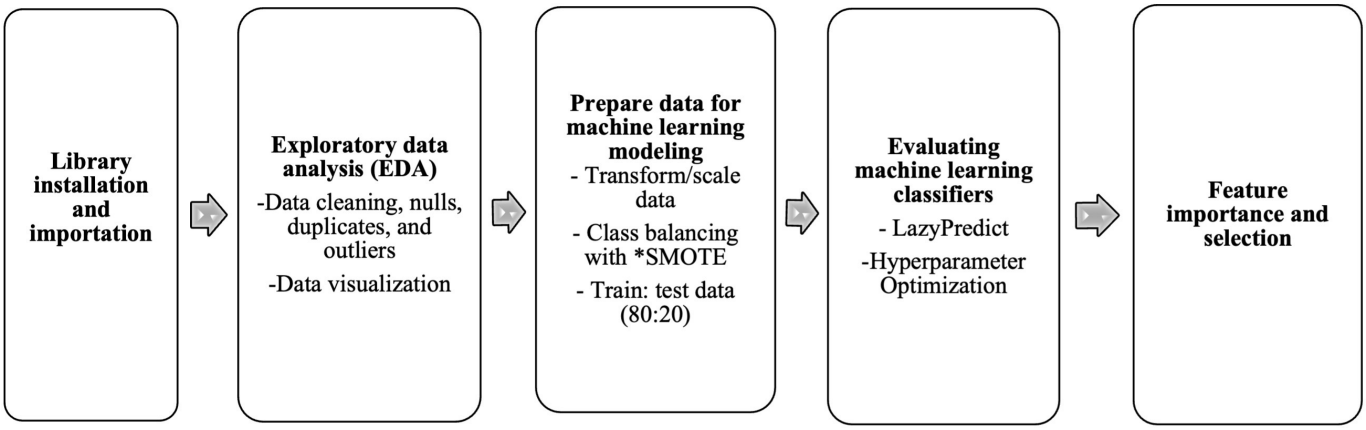
## Correlation Analysis

Data cleaning, exploratory analysis, and feature engineering were performed in Google Colab (with Python 3.10; Python Software Foundation). The target variable was specified as "blood glucose," where 1 indicated an RBG concentration ≥140 mg/dl and 0 indicated an RBG concentration <140 mg/dl. Independent variables included age (integer), sex (integer), BMI (float), smoking status (integer), ECG (float), hemoglobin (float), cholesterol (float), uric acid (float), SBP (integer), DBP (integer), normal BP (integer), elevated BP (integer), prehypertension (integer), hypertension (integer), ISH (integer), IDH (integer), prediabetes (integer), diabetes (integer), normal glucose (integer), abnormal ECG values (integer), and normal ECG values (integer). The dataset was checked and visualized for missingness using seaborn heatmap (Figure S1 in Multimedia Appendix 1). Missing values were replaced with column mean (for continuous variables) or mode (for categorical variables). Duplicate rows and outliers were dropped before encoding categorical variables and creating dummy variables. Subsequently, we created a heatmap for the correlation of independent variables with the target column in descending order. The cleaned dataset was then scaled for subsequent training of machine learning models. A $P$ value ≤.05 was considered statistically significant.

## Machine Learning Algorithms and Evaluation

The study adopted 26 supervised classification algorithms and compared their accuracies to identify the best-performing model for predicting high blood glucose, which was defined in this study as an RBG concentration ≥140 mg/dl (Figure 1). Specifically, after the installation and importation of Sci-Kit Learn libraries [22], we carried out data cleaning, exploration, and scaling to improve the efficiency of our model (Multimedia Appendix 1). Imbalances in the distribution of hyperglycemia cases and noncases within the dataset might affect the model's performance. Addressing this imbalance and validating the model on balanced datasets could enhance its robustness. To address the class imbalance in the outcome variable (blood glucose level), we adopted the synthetic minority oversampling technique (SMOTE). SMOTE tackled the underrepresentation of the minority class and rebalanced the class distribution for equitability [23]. After resampling, we split the data into training and test sets at a ratio of 80:20, respectively, using the train_test_split function in Sci-Kit Learn. We went further to select and rank the performances of the machine learning algorithms using LazyPredict to obtain the weighted average of the $F_1$-scores and accuracy scores as well as the receiver operating characteristic–area under the curve (ROC-AUC) score. For hyperparameter optimization, we adopted GridSearchCV [24]. The grid search technique constructs many versions of the model with all possible combinations of hyperparameters to return the best one [25]. Subsequently, we determined feature importance to provide insight into which features are most associated with elevated blood glucose levels using the best-performing model. To operationalize the best-performing model generated at scale, the training file was stored as a serialized pickle file. Subsequently, we used the Fast Application Programming Interface in Google Colab [26] to make an inference call from the model using the predict() function and generated our application programming interface. Pyngrok was used to open secure tunnels from public URLs to the local host.

**Figure 1.** Pipeline for model development. SMOTE: synthetic minority oversampling technique.

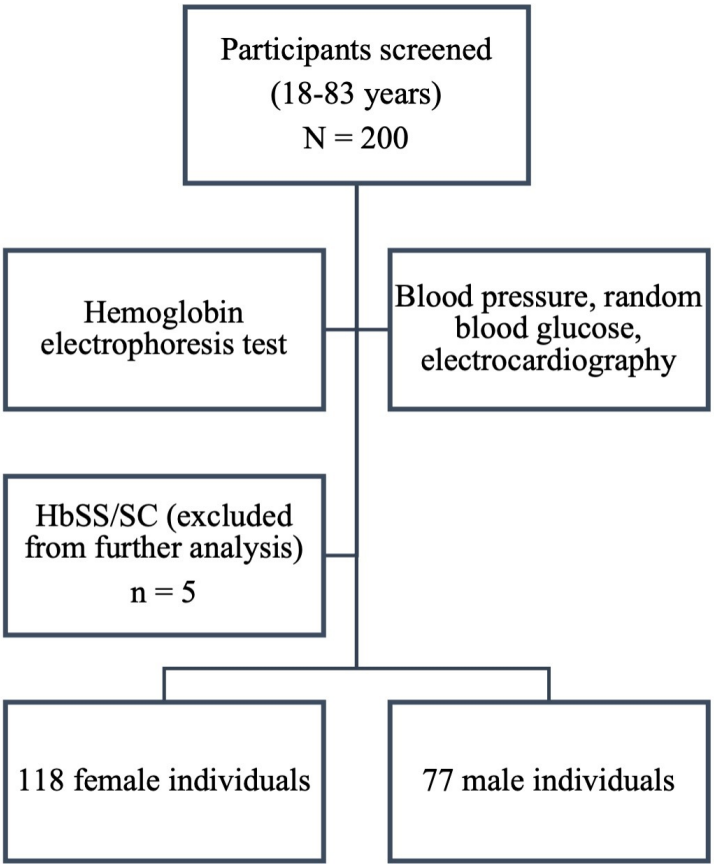



# Results

## Cohort Description

A total of 200 participants ages 18-83 years were enrolled in the cohort. However, after hemoglobin electrophoresis screening, 5 participants were found to possess the hemoglobins SS and SC genotypes and were excluded from further analysis. A total of 118 female and 77 male participants were included (Figure 2 and Figure S2 in Multimedia Appendix 1).

**Figure 2.** Participant recruitment and screening.




## Correlation Analysis

Participants were categorized into six age groups: 18-29, 30-39, 30-49, 50-59, 60-69, and ≥70 years. Of the 195 participants, normal BP, elevated BP, and prehypertension were recorded in 63 (32.3%), 8 (4.1%), and 18 (9.2%) participants, respectively. A total of 41 (21%) participants presented with hypertension, of which 34 (83%) were female (Figure S3 in Multimedia Appendix 1). Age-based analysis revealed an inverse relationship between normotension and age ($r$=–0.88; $P$=.02). Consistently, hypertension increased with age ($r$=0.53; $P$=.27), peaking between 50-59 years (Figure 3). Of the 195 participants, ISH and IDH were recorded in 16 (8.2%) and 15 (7.7%) participants, respectively, with female participants recording a higher prevalence of ISH (11/16, 69%) and male participants reporting a higher prevalence of IDH (11/15, 73%; Figure S4 in Multimedia Appendix 1). There was a positive correlation between ISH and participants' age ($r$=0.86; $P$=.03), whereas IDH was inversely correlated with age ($r$=–0.71; $P$=.11; Figure 4). We went further to examine the heart rates of the participants and observed an age-dependent increase in the percentage of participants with abnormal ECG values peaking between ages 60-69 years (Figure 5). However, no significant difference was observed in the ECG values of male and female participants ($\sqrt{X^2}$=0.13; $P$=.72; Figure S5 in Multimedia Appendix 1). An RBG value between 140-199 mg/dl (prediabetes) was detected in 22 (11.3%) and diabetes was suspected in 5 (2.6%) of the 195 participants. A total of 163 (85.8%) participants had normal blood glucose. Though not statistically significant, an inverse relationship ($r$=–0.81; $P$=.06) was observed between age and normal glucose level, and the frequency of prediabetes ($r$=0.63; $P$=.19) and suspected diabetes ($r$=0.58; $P$=.24) seemed to increase with age (Figure S6 in Multimedia Appendix 1). Meanwhile, a correlation matrix between each independent variable and the target column (blood glucose level) showed that age had the highest ranking even though the correlation coefficient was weak (Figure 6 and Figure S7 in Multimedia Appendix 1).

**Figure 3.** Age-based analysis of BP. Percentage of participants with normal BP reduced with increases in age ($r$=–0.88; $P$=.02). Prevalence of HTN increased with age ($r$=0.53; $P$=.27), peaking between 50-59 years. BP: blood pressure; HTN: hypertension.

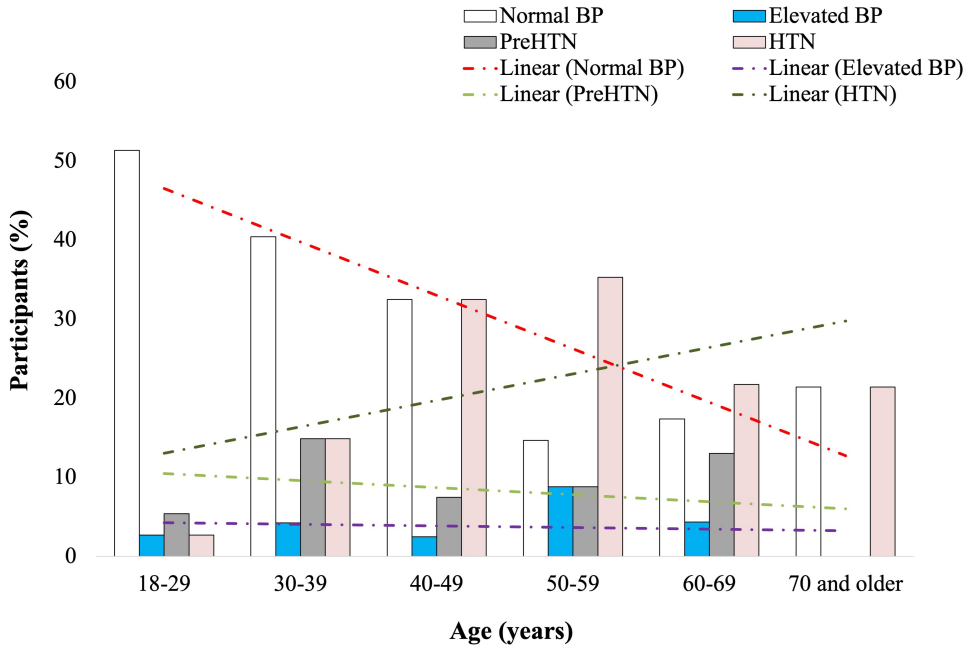



**Figure 4.** Age-based analysis of ISH and IDH. ISH increased with participants' age ($r$=0.86; $P$=.03), unlike IDH ($r$=–0.71; $P$=.11). IDH: isolated diastolic hypertension; ISH: isolated systolic hypertension.

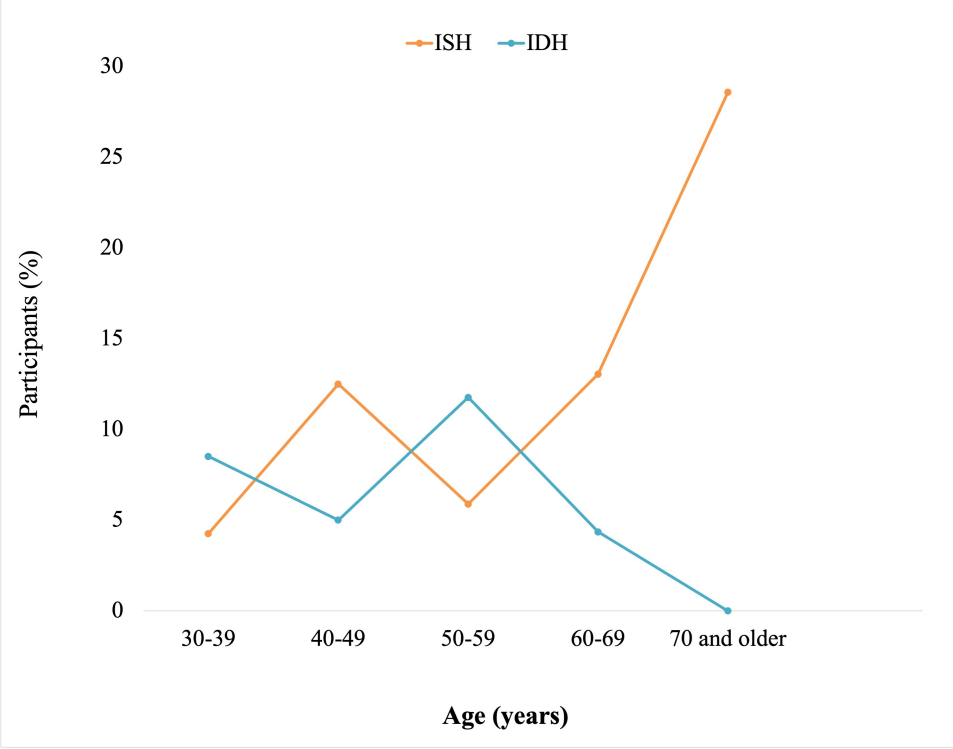


**Figure 5.** Age-based ECG analysis. Age-dependent increase in the percentage of participants with abnormal ECG values peaking between ages 60-69 years. ECG: electrocardiogram.

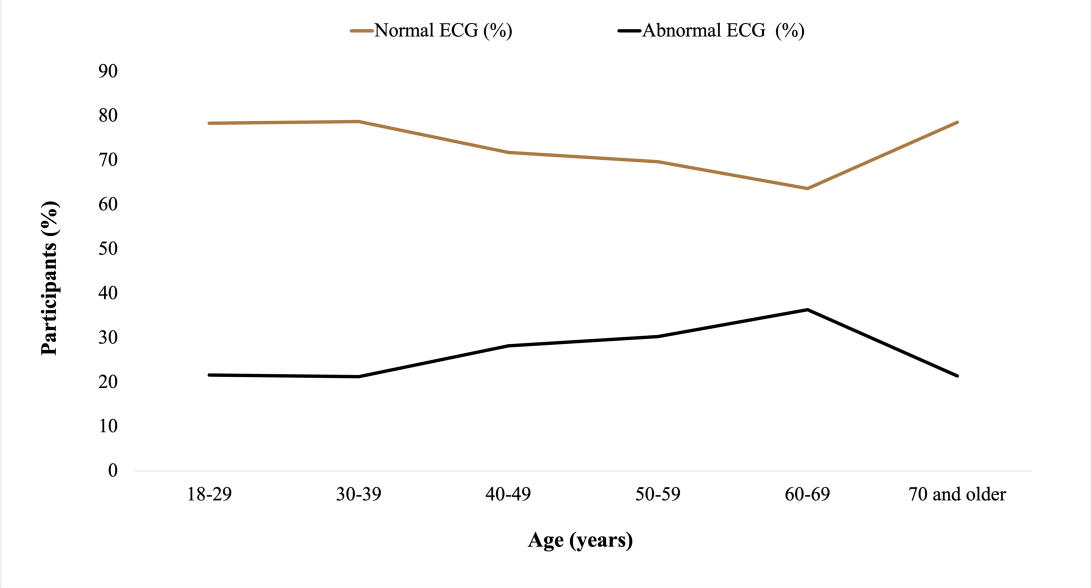
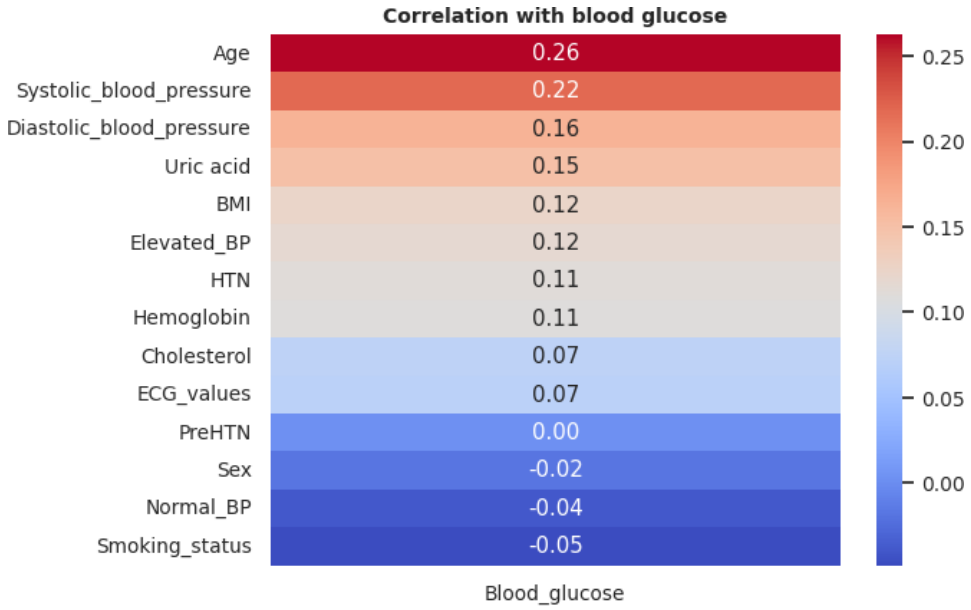



**Figure 6.** Correlation matrix of independent variables with the outcome variable. BP: blood pressure; ECG: electrocardiogram; HTN: hypertension.




## Machine Learning Algorithms and Evaluation

Following data cleaning, transformation (Figure S8 in Multimedia Appendix 1), and observation of a class imbalance in the target variable (Figure S9 in Multimedia Appendix 1), whereby the raw dataset demonstrated that 163 (83.6%) of the 195 participants had normal blood glucose (0), while 32 (16.4%) had a high blood glucose level (1), rebalancing was established with SMOTE to yield an even representation of both categories of blood glucose level (counter: 0: 163; 1: 163). When the performance of each classifier was tested, the reports showed that the random forest classifier (Figures 7 and 8) gave the best accuracy (accuracy score 0.89; ROC-AUC score 0.89; $F_1$-score 0.89), followed by extra trees (accuracy score 0.88; ROC-AUC score 0.88; $F_1$-score=0.88) and extreme gradient boosting classifiers (accuracy score 0.86; ROC-AUC score 0.86; $F_1$-score 0.86; Figure 7B and Table S2 in Multimedia Appendix 1).

To determine the importance of each variable (feature) to the outcome (blood glucose level), we carried out a random forest feature analysis. The importance of a feature is calculated based on how much the tree nodes that use that feature reduce impurity across all trees in the forest. The key findings showed that uric acid and age were the most important features associated with elevated blood glucose (Table 1), followed by SBP and BMI.

**Figure 7.** Accuracy scores of machine learning classifiers (A) before class rebalancing with the synthetic minority oversampling technique and (B) after class rebalancing with the synthetic minority oversampling technique. CV: cross-validation; LGBM: light gradient boosting machine; NB: naive Bayes; SGD: stochastic gradient descent; SVC: support vector classification; XGB: extreme gradient boosting.

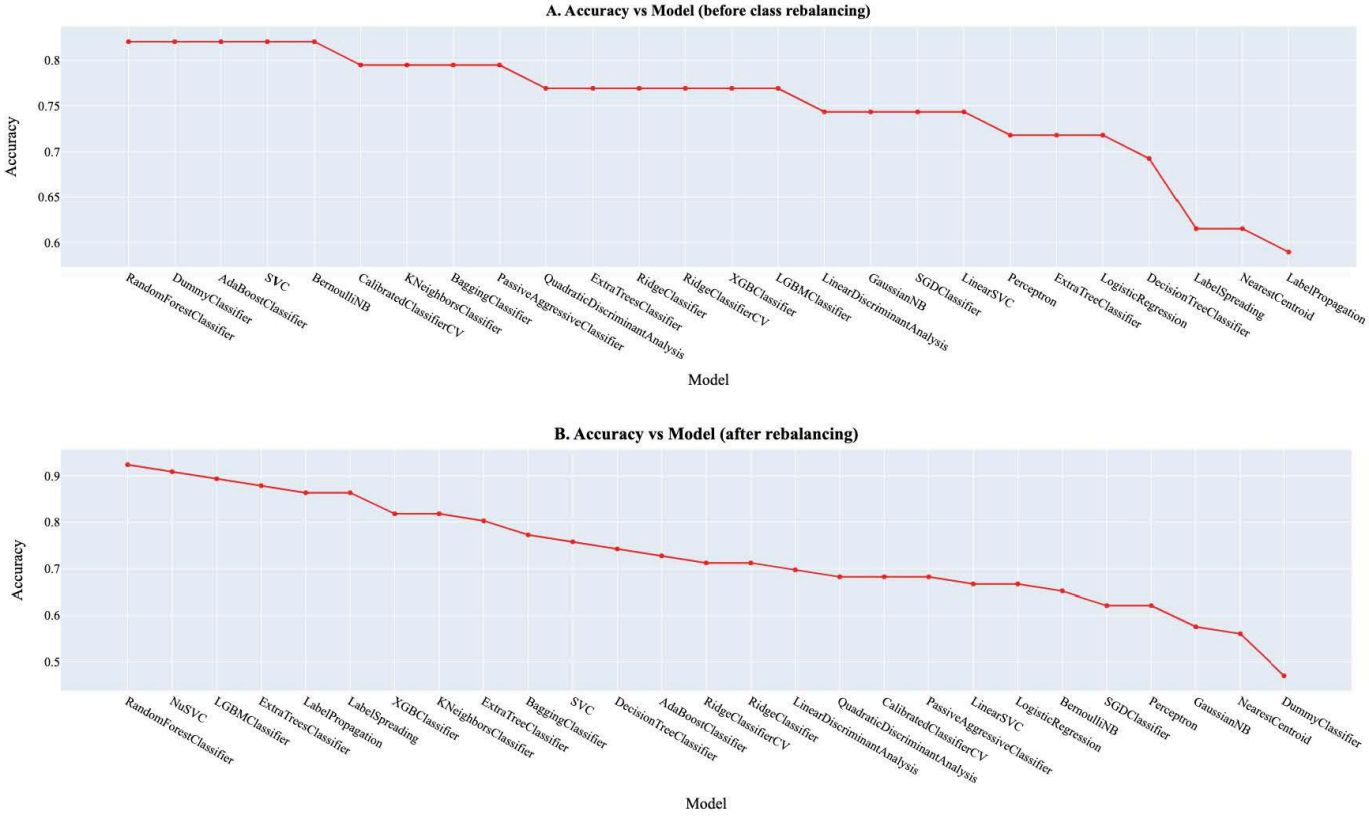
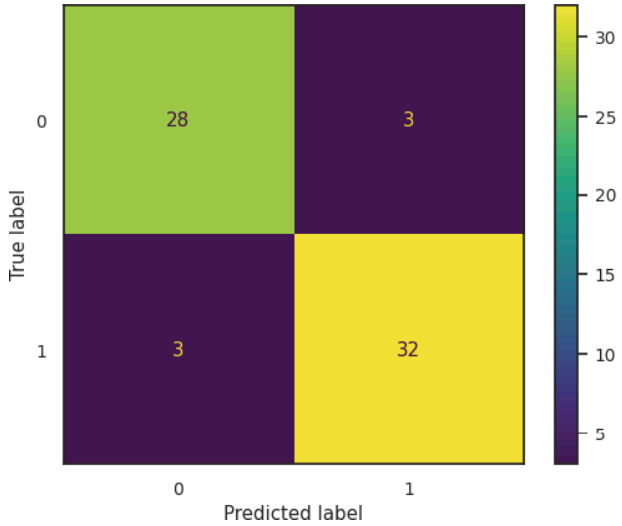



**Figure 8.** Random forest confusion matrix showing a visual representation of the true vs predicted labels. True positive: the values that were positive and were predicted positive, that is, 31 cases of hyperglycemia were predicted correctly by the model. False positive: the values that were negative but falsely predicted as positive. In this case, only 3 cases were false positives. False negative: the values that were positive but falsely predicted as negative. In this instance, there were 4 false negatives. True negative: the values that were negative and were predicted negative. Here, 28 cases were detected. In all, the weighted average of the accuracy score and $F_1$-score were 0.89 and 0.89, respectively. Precision is a metric that quantifies the accuracy of a classifier by determining the number of correctly identified members of a class divided by all instances where the model predicted that specific class. In the context of hyperglycemia prediction, precision would be the count of accurate predictions of hyperglycemia divided by the total instances where the classifier predicted "hyperglycemia," regardless of correctness. Recall, on the other hand, measures the effectiveness of a classifier in correctly identifying members of a class by dividing the number of correctly identified instances by the total number of actual members in that class. In the hyperglycemia scenario, recall would represent the number of actual participants with hyperglycemia correctly identified by the classifier. The $F_1$-score is a composite metric that combines both precision and recall into a single value. It provides a concise evaluation of a classifier's performance. A high $F_1$-score indicates that both precision and recall are high, while a low $F_1$-score suggests that one or both metrics are low. This metric is particularly useful for quickly assessing whether a classifier effectively identifies members of a class or if it resorts to shortcuts, such as indiscriminately classifying everything as a member of a larger class. avg: average.




Classification report

| | Precision | Recall | $F_1$-score | Support |
|---|---|---|---|---|
| 0 | 0.90 | 0.90 | 0.90 | 31 |
| 1 | 0.91 | 0.91 | 0.91 | 35 |
| Accuracy | | | 0.91 | 66 |
| Macro avg | 0.91 | 0.91 | 0.91 | 66 |
| Weighted avg | 0.91 | 0.91 | 0.91 | 66 |

**Table 1.** Blood glucose predictors.

| Feature | Importance[a] |
|---|---|
| Age | 0.17 |
| Uric acid | 0.17 |
| BMI | 0.12 |
| Systolic blood pressure | 0.11 |
| Diastolic blood pressure | 0.10 |
| Cholesterol | 0.09 |
| Hemoglobin | 0.08 |
| Electrocardiogram values | 0.04 |
| Sex | 0.04 |
| Normal blood pressure | 0.03 |
| Elevated blood pressure | 0.02 |
| Hypertension | 0.02 |
| Pre-hypertension | 0.01 |
| Smoking status | 0.01 |

[a]The importance of a feature is calculated based on how much the tree nodes that use that feature reduce impurity across all trees in the forest.

# Discussion

## Principal Findings

NCDs, such as cancer, cardiovascular diseases, and diabetes, are progressively becoming the primary causes of mortality in sub-Saharan Africa [27]. This epidemiological shift is primarily attributed to limitations in implementing crucial control measures, such as prevention and early detection [1]. This research focused on exploring key clinical indices of NCDs in individuals who are asymptomatic. The application of machine learning in disease prediction is now well established for its immense potential in analyzing complex datasets and uncovering patterns that may elude human detection [28-31]. This investigation employed various machine learning algorithms to predict hyperglycemia to enable early identification of individuals at risk of developing diabetes. The study identified suspected hypertension in 21% of study participants, underscoring the urgency of addressing hypertension as a major health challenge in the country. Furthermore, a notable increase in the prevalence of hypertension with advancing age was observed. However, the investigation into hypertension subtypes revealed a dual phenomenon: a pronounced increase in systolic hypertension with age and a concomitant reduction in diastolic hypertension.

Several factors may contribute to the observed age-related increase in systolic hypertension. Physiological changes, alterations in vascular reactivity, and lifestyle factors could play decisive roles in driving the upward trajectory of SBP with advancing age [32,33]. In contrast, the age-related reduction in diastolic hypertension may be associated with changes in arterial compliance, heart rate dynamics, or other physiological adaptations over the aging process [34]. Recognizing these dual dynamics holds significant clinical implications, necessitating tailored screening protocols and interventions to address the unique challenges posed by hypertension in different age groups.

Moreover, a sex disparity was observed, with systolic hypertension being more prevalent in female participants and diastolic hypertension being more common in male participants. This sex difference may be linked to heart rate variability or hormonal influences, particularly fluctuations in estrogen levels in female individuals. However, understanding how blood vessels respond to changes in pressure and the potential impact on SBP would be crucial in deciphering these sex disparities [35-37]. Therefore, tailoring screening protocols and interventions to address the unique challenges posed by hypertension in different age groups and sexes is essential to mitigate the overall burden of this condition.

ECG is a pivotal tool for assessing cardiac health, and its interpretation can provide valuable insights into cardiovascular conditions. Our investigation revealed a remarkable age-dependent pattern in abnormal ECG values, reaching a peak at 70 years. Advancing age often coincides with a myriad of physiological changes, including alterations in cardiac structure and function [38-40]. A comprehensive exploration of these factors is essential for delineating the intricate relationship between aging and abnormal ECG findings.

The global burden of diabetes is well-documented [41-43], but our investigation into supposedly healthy individuals has unearthed a concerning revelation. Despite outward appearances of health, there existed a relatively high prevalence of suspected prediabetes and diabetes in the cohort. This underscores the importance of probing beyond outward health markers to understand the latent metabolic landscape [44-47]. This prompts a reevaluation of health screening protocols to incorporate metabolic parameters in apparently healthy populations. Early detection and intervention strategies should be tailored to encompass metabolic assessments,

providing an opportunity for targeted preventive measures and lifestyle modifications.

In the realm of predictive modeling, selecting the most effective machine learning algorithm is paramount. Our study, aimed at evaluating various algorithms, revealed insightful findings regarding their predictive performances. Upon meticulous evaluation, random forest emerged as the top-performing algorithm, consistently delivering the highest accuracy among the tested models. The success of the random forest algorithm can be attributed to its ensemble learning nature [48,49], which harnesses the collective power of multiple decision trees. This enables robustness against overfitting, enhanced generalization, and effective handling of complex datasets with diverse features. The observed superiority of random forest in our study has profound implications for future applications, suggesting its applicability across diverse datasets and underscoring its potential as a reliable choice for achieving high predictive accuracy.

To investigate the intricate determinants of hyperglycemia, our study employed a robust feature importance analysis, with compelling results showcasing uric acid and age as the most influential predictors. Uric acid's prominence as a predictor of hyperglycemia adds a unique dimension to our understanding of metabolic health. While traditionally associated with conditions like gout, our findings suggest a potential link between hyperuricemia and hyperglycemia, urging further exploration into the underlying physiological mechanisms. The identification of age as a key predictor aligns with existing knowledge regarding the age-associated risk of hyperglycemia [49-51]. Our findings reinforce the significance of age as a robust indicator, reflecting the cumulative impact of aging processes on metabolic health and glucose regulation. The recognition of uric acid and age as pivotal predictors holds significant clinical implications. Health care practitioners can leverage these findings to enhance risk assessment strategies for hyperglycemia. Incorporating uric acid measurements and age considerations into routine screenings may facilitate early identification of individuals at heightened risk, enabling proactive interventions. While our study sheds light on the importance of uric acid and age, further research is warranted to unravel the intricate relationships and mechanisms underlying these associations. Longitudinal studies exploring the dynamic interplay between uric acid, age, and hyperglycemia can deepen our understanding and inform targeted interventions.

## Limitations and Future Direction

While our study provides valuable insights into predicting hyperglycemia using machine learning in undiagnosed individuals, it is essential to acknowledge certain limitations that may impact interpretation. First, the size of our cohort may limit the generalizability of the results. A larger and more diverse sample could enhance the external validity of the predictive model. Furthermore, the study did not account for potential variations in clinical practice, including differences in diagnostic criteria. For instance, the study did not take into consideration orthostatic hypotension, a decrease in SBP ≥20 mmHg or a DBP decrease of ≥10 mmHg within 3 minutes of standing, especially in older individuals [19]. Although seats were provided to participants, we could not accurately document how long participants had been standing before attending the screening. Besides, phenomena such as postprandial hypotension (a reduction in BP after meals, a common cause of syncope and falls in older individuals who are healthy and have hypertension), circadian BP variability, and white-coat (nonsustained) hypertension, especially in older adults, were not factored into the analyses [52-54]. As such, incorporating standardized criteria across diverse health care settings could enhance our model's clinical applicability.

Moreover, the study did not dissect the influence of ethnicity and genetics on hyperglycemia [55,56]. Future research could explore these aspects to provide a more comprehensive understanding of predictive factors. Since the dataset primarily comprises information from a specific geographic location or demographic group, extrapolating the findings to other populations requires caution as regional variations in lifestyle, genetics, and health care practices may influence the performance of the predictive model. In addition, the cross-sectional nature of our study limits our ability to establish causation or assess changes over time. Therefore, longitudinal studies would be beneficial to understand the dynamic nature of hyperglycemia predictors. The model's performance was evaluated on the same dataset used for training, raising the potential for overfitting. External validation on an independent dataset would be crucial to assess its generalizability and reliability in real-world scenarios. Lastly, the importance of a feature in a random forest model does not necessarily mean a causal relationship and other models might find different results if additional features are introduced. Future approaches are expected to accommodate more features and larger datasets. This will account for the deployment of built and containerized models as publicly accessible web applications. Nevertheless, this study has expounded the potential of machine learning for early disease detection, risk assessment strategies, proactive interventions, and targeted therapeutic design.

## Conclusions

This study has made a substantial contribution to the expanding domain of predictive modeling and offers promising implications for enhancing early detection and personalized risk assessment, particularly in the context of hyperglycemia and its potential association with diabetes. The research has not only brought to light the prevalence of undiagnosed hypertension and isolated systolic and diastolic hypertension but also highlighted factors associated with elevated blood glucose within the population. The findings of this study emphasize the significance of regular screening, effective intervention strategies, and targeted therapeutic designs. Collectively, the results contribute to the overarching effort to enhance health care outcomes through proactive and tailored approaches.

**Acknowledgments**

The authors appreciate the study participants and Ijede community leaders for their cooperation during the screening exercise. KO was supported by the Fogarty Emerging Global Leader Grant (NIH-K43TW011926) from the US National Institutes of Health and an APTI-18-07 grant from the African Academy of Sciences in partnership with the Bill and Melinda Gates Foundation. The funders had no role in the study design, data collection, analysis, decision to publish, or preparation of the manuscript.

## Data Availability

The datasets supporting the conclusions of this paper are within the manuscript and Multimedia Appendices 1 and 2. The Google Colab Python script used for data analysis and machine learning has been deposited in our GitHub page [24].

## Authors' Contributions

KO conceived and designed the study. KO, FL, AO, BE, YA, and OA implemented the field study. KO, FL, and BE carried out the laboratory experiments. KO carried out the data analysis, including machine learning. KO drafted the manuscript. KO, OA, and BS edited the manuscript. All authors read and approved the final manuscript.

## Conflicts of Interest

None declared.

## Multimedia Appendix 1

Supplementary methods and figures.
[DOCX File (Microsoft Word File), 987 KB-Multimedia Appendix 1]

## Multimedia Appendix 2

Full dataset.
[XLS File (Microsoft Excel File), 60 KB-Multimedia Appendix 2]

## References

1. Bigna JJ, Noubiap JJ. The rising burden of non-communicable diseases in sub-Saharan Africa. Lancet Glob Health. Oct 2019;7(10):e1295-e1296. [doi: 10.1016/S2214-109X(19)30370-5] [Medline: 31537347]
2. Cross SH, Mehra MR, Bhatt DL, et al. Rural-urban differences in cardiovascular mortality in the US, 1999-2017. JAMA. May 12, 2020;323(18):1852-1854. [doi: 10.1001/jama.2020.2047] [Medline: 32396176]
3. Turecamo SE, Xu M, Dixon D, et al. Association of rurality with risk of heart failure. JAMA Cardiol. Mar 1, 2023;8(3):231-239. [doi: 10.1001/jamacardio.2022.5211] [Medline: 36696094]
4. Khayat S, Dolatian M, Navidian A, Mahmoodi Z, Sharifi N, Kasaeian A. Lifestyles in suburban populations: a systematic review. Electron Physician. Jul 25, 2017;9(7):4791-4800. [doi: 10.19082/4791] [Medline: 28894537]
5. Kolié D, Van De Pas R, Codjia L, Zurn P. Increasing the availability of health workers in rural sub-Saharan Africa: a scoping review of rural pipeline programmes. Hum Resour Health. Mar 14, 2023;21(1):20. [doi: 10.1186/s12960-023-00801-z] [Medline: 36918864]
6. Ngene NC, Khaliq OP, Moodley J. Inequality in health care services in urban and rural settings in South Africa. Afr J Reprod Health. May 2023;27(5s):87-95. [doi: 10.29063/ajrh2023/v27i5s.11] [Medline: 37584924]
7. Jane Ling MY, Ahmad N, Aizuddin AN. Risk perception of non-communicable diseases: a systematic review on its assessment and associated factors. PLoS One. Jun 1, 2023;18(6):e0286518. [doi: 10.1371/journal.pone.0286518] [Medline: 37262079]
8. Tohidinezhad F, Khorsand A, Zakavi SR, et al. The burden and predisposing factors of non-communicable diseases in Mashhad University of Medical Sciences personnel: a prospective 15-year organizational cohort study protocol and baseline assessment. BMC Public Health. Nov 2, 2020;20(1):1637. [doi: 10.1186/s12889-020-09704-3] [Medline: 33138802]
9. Alanazi R. Identification and prediction of chronic diseases using machine learning approach. J Healthc Eng. Feb 25, 2022;2022:2826127. [doi: 10.1155/2022/2826127] [Medline: 35251563]
10. Park DJ, Park MW, Lee H, Kim YJ, Kim Y, Park YH. Development of machine learning model for diagnostic disease prediction based on laboratory tests. Sci Rep. Apr 7, 2021;11(1):7567. [doi: 10.1038/s41598-021-87171-5] [Medline: 33828178]
11. Uddin S, Khan A, Hossain ME, Moni MA. Comparing different supervised machine learning algorithms for disease prediction. BMC Med Inform Decis Mak. Dec 21, 2019;19(1):281. [doi: 10.1186/s12911-019-1004-8] [Medline: 31864346]
12. Wang M, Ge W, Apthorp D, Suominen H. Robust feature engineering for Parkinson disease diagnosis: new machine learning techniques. JMIR Biomed Eng. Jul 27, 2020;5(1):e13611. [doi: 10.2196/13611]

13. Sampa MB, Biswas T, Rahman MS, Aziz NHBA, Hossain MN, Aziz NAA. A machine learning web app to predict diabetic blood glucose based on a basic noninvasive health checkup, sociodemographic characteristics, and dietary information: case study. JMIR Diabetes. Nov 24, 2023;8:e49113. [doi: 10.2196/49113] [Medline: 37999944]
14. Sampa MB, Hossain MN, Hoque MR, et al. Blood uric acid prediction with machine learning: model development and performance comparison. JMIR Med Inform. Oct 8, 2020;8(10):e18331. [doi: 10.2196/18331] [Medline: 33030442]
15. Abd El-Hafeez T, Shams MY, Elshaier YAMM, Farghaly HM, Hassanien AE. Harnessing machine learning to find synergistic combinations for FDA-approved cancer drugs. Sci Rep. Jan 29, 2024;14(1):2428. [doi: 10.1038/s41598-024-52814-w] [Medline: 38287066]
16. Hassan E, Abd El-Hafeez T, Shams MY. Optimizing classification of diseases through language model analysis of symptoms. Sci Rep. Jan 17, 2024;14(1):1507. [doi: 10.1038/s41598-024-51615-5] [Medline: 38233458]
17. Keohane EM, Otto CN, Walenga JM. Rodak's Hematology, 6th Edition. Elsevier Health Sciences; 2015. ISBN: 9780323936507
18. Yousefi M, Najafi Saleh H, Yaseri M, Jalilzadeh M, Mohammadi AA. Association of consumption of excess hard water, body mass index and waist circumference with risk of hypertension in individuals living in hard and soft water areas. Environ Geochem Health. Jun 2019;41(3):1213-1221. [doi: 10.1007/s10653-018-0206-9] [Medline: 30390219]
19. Tan JL, Thakur K. Systolic Hypertension. StatPearls Publishing; 2023. URL: https://www.ncbi.nlm.nih.gov/books/NBK482472/ [Accessed 2024-08-07]
20. Whelton PK, Carey RM, Aronow WS, et al. 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APHA/ASH/ASPC/NMA/PCNA guideline for the prevention, detection, evaluation, and management of high blood pressure in adults: executive summary: a report of the American College of Cardiology/American Heart Association task force on clinical practice guidelines. Circulation. Oct 23, 2018;138(17):e426-e483. [doi: 10.1161/CIR.0000000000000597] [Medline: 30354655]
21. American Diabetes Association. Diagnosis and classification of diabetes mellitus. Diabetes Care. Jan 2010;33(Suppl 1):S62-S69. [doi: 10.2337/dc10-S062] [Medline: 20042775]
22. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. J Machine Learning Res. Oct 2011;12(2011):2825-2830. URL: https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf [Accessed 2024-08-07]
23. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell. Jun 1, 2002;16:321-357. [doi: 10.1613/jair.953]
24. Oyebola K. Machine learning prediction of elevated blood glucose in a cohort of apparently healthy adults. GitHub. URL: https://github.com/oyebolakolapo/Machine-Learning-Prediction-of-Elevated-Blood-Glucose-in-a-Cohort-of-Apparently-Healthy-Adults [Accessed 2024-08-07]
25. Buyya R, Hernandez SM, Kovvur RMR, Sarma TH, editors. Computational Intelligence and Data Analytics: Proceedings of ICCIDA 2022. Springer; 2022. [doi: 10.1007/978-981-19-3391-2]
26. Lathkar M. High-Performance Web Apps with FastAPI: The Asynchronous Web Framework Based on Modern Python. Apress; 2023. [doi: 10.1007/978-1-4842-9178-8]
27. Katende D, Kasamba I, Sekitoleko I, et al. Medium-to-long term sustainability of a health systems intervention to improve service readiness and quality of non-communicable disease (NCD) patient care and experience at primary care settings in Uganda. BMC Health Serv Res. Sep 22, 2023;23(1):1022. [doi: 10.1186/s12913-023-09983-7] [Medline: 37737179]
28. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthc J. Jun 2019;6(2):94-98. [doi: 10.7861/futurehosp.6-2-94] [Medline: 31363513]
29. Abdel Hady DA, Abd El-Hafeez T. Predicting female pelvic tilt and lumbar angle using machine learning in case of urinary incontinence and sexual dysfunction. Sci Rep. Oct 20, 2023;13(1):17940. [doi: 10.1038/s41598-023-44964-0] [Medline: 37863988]
30. Eliwa EHI, El Koshiry AM, Abd El-Hafeez T, Farghaly HM. Utilizing convolutional neural networks to classify monkeypox skin lesions. Sci Rep. Sep 3, 2023;13(1):14495. [doi: 10.1038/s41598-023-41545-z] [Medline: 37661211]
31. Mamdouh Farghaly H, Shams MY, Abd El-Hafeez T. Hepatitis C virus prediction based on machine learning framework: a real-world case study in Egypt. Knowledge Inf Syst. Jun 2023;65(6):2595-2617. [doi: 10.1007/s10115-023-01851-4]
32. Sharifi-Rad J, Rodrigues CF, Sharopov F, et al. Diet, lifestyle and cardiovascular diseases: linking pathophysiology to cardioprotective effects of natural bioactive compounds. Int J Environ Res Public Health. Mar 30, 2020;17(7):2326. [doi: 10.3390/ijerph17072326] [Medline: 32235611]
33. Liu R, Li D, Yang Y, Hu Y, Wu S, Tian Y. Systolic blood pressure trajectories and the progression of arterial stiffness in Chinese adults. Int J Environ Res Public Health. Aug 15, 2022;19(16):10046. [doi: 10.3390/ijerph191610046] [Medline: 36011682]

34. Singh JN, Nguyen T, Kerndt CC, Dhamoon AS. Physiology, Blood Pressure Age Related Changes. StatPearls Publishing; 2023. [Medline: 30725982]
35. Song JJ, Ma Z, Wang J, Chen LX, Zhong JC. Gender differences in hypertension. J Cardiovasc Transl Res. Feb 2020;13(1):47-54. [doi: 10.1007/s12265-019-09888-z] [Medline: 31044374]
36. Wu J, Jiao B, Fan Y. Urbanization and systolic/diastolic blood pressure from a gender perspective: separating longitudinal from cross-sectional association. Health Place. May 2022;75:102778. [doi: 10.1016/j.healthplace.2022.102778] [Medline: 35339955]
37. Midtbø H, Gerdts E. Sex disparities in blood pressure development: time for action. Eur J Prev Cardiol. Feb 19, 2022;29(1):178-179. [doi: 10.1093/eurjpc/zwab109] [Medline: 34223618]
38. Fleg JL, Forman DE, et al. Aging changes in cardiovascular structure and function. In: Waldstein SR, Kop WJ, Suarez ED, editors. Handbook of Cardiovascular Behavioral Medicine. Springer; 2022:127-162. [doi: 10.1007/978-0-387-85960-6_6]
39. Fleg JL, Strait J. Age-associated changes in cardiovascular structure and function: a fertile milieu for future disease. Heart Fail Rev. Sep 2012;17(4-5):545-554. [doi: 10.1007/s10741-011-9270-2] [Medline: 21809160]
40. Hacker TA, McKiernan SH, Douglas PS, Wanagat J, Aiken JM. Age-related changes in cardiac structure and function in Fischer 344 x Brown Norway hybrid rats. Am J Physiol Heart Circ Physiol. Jan 2006;290(1):H304-H311. [doi: 10.1152/ajpheart.00290.2005] [Medline: 16143657]
41. King H, Aubert RE, Herman WH. Global burden of diabetes, 1995-2025: prevalence, numerical estimates, and projections. Diabetes Care. Sep 1998;21(9):1414-1431. [doi: 10.2337/diacare.21.9.1414] [Medline: 9727886]
42. Herman WH. The global burden of diabetes: an overview. In: Dagogo-Jack S, editor. Diabetes Mellitus in Developing Countries and Underserved Communities. Springer; 2017:1-5. [doi: 10.1007/978-3-319-41559-8_1]
43. Ong KL, Stafford LK, McLaughlin SA. Global, regional, and national burden of diabetes from 1990 to 2021, with projections of prevalence to 2050: a systematic analysis for the Global Burden of Disease Study 2021. Lancet. Jul 15, 2023;402(10397):203-234. [doi: 10.1016/S0140-6736(23)01301-6] [Medline: 37356446]
44. Huang PL. A comprehensive definition for metabolic syndrome. Dis Model Mech. 2009;2(5-6):231-237. [doi: 10.1242/dmm.001180] [Medline: 19407331]
45. Rafaqat S, Sharif S, Majeed M, Naz S, Manzoor F, Rafaqat S. Biomarkers of metabolic syndrome: role in pathogenesis and pathophysiology of atrial fibrillation. J Atr Fibrillation. Aug 31, 2021;14(2):20200495. [doi: 10.4022/jafib.20200495] [Medline: 34950373]
46. Srikanthan K, Feyh A, Visweshwar H, Shapiro JI, Sodhi K. Systematic review of metabolic syndrome biomarkers: a panel for early detection, management, and risk stratification in the West Virginian population. Int J Med Sci. Jan 1, 2016;13(1):25-38. [doi: 10.7150/ijms.13800] [Medline: 26816492]
47. Madhusoodanan J. Searching for better biomarkers for metabolic syndrome. ACS Cent Sci. Jun 22, 2022;8(6):682-685. [doi: 10.1021/acscentsci.2c00629] [Medline: 35756383]
48. Schonlau M, Zou RY. The random forest algorithm for statistical learning. Stata J. Mar 24, 2020;20(1):3-29. [doi: 10.1177/1536867X20909688]
49. Ghaffar Nia N, Kaplanoglu E, Nasab A. Evaluation of artificial intelligence techniques in disease diagnosis and prediction. Discov Artif Intell. Jan 30, 2023;3(1). [doi: 10.1007/s44163-023-00049-5]
50. Longo M, Bellastella G, Maiorino MI, Meier JJ, Esposito K, Giugliano D. Diabetes and aging: from treatment goals to pharmacologic therapy. Front Endocrinol (Lausanne). Feb 18, 2019;10:45. [doi: 10.3389/fendo.2019.00045] [Medline: 30833929]
51. Yan Z, Cai M, Han X, Chen Q, Lu H. The interaction between age and risk factors for diabetes and prediabetes: a community-based cross-sectional study. Diabetes Metab Syndr Obes. Jan 11, 2023;16:85-93. [doi: 10.2147/DMSO.S390857] [Medline: 36760587]
52. Nuredini G, Saunders A, Rajkumar C, Okorie M. Current status of white coat hypertension: where are we? Ther Adv Cardiovasc Dis. 2020;14:1753944720931637. [doi: 10.1177/1753944720931637] [Medline: 32580646]
53. Franklin SS, Thijs L, Hansen TW, O'Brien E, Staessen JA. White-coat hypertension. Hypertension. Sep 16, 2013;62(6):982-987. [doi: 10.1161/HYPERTENSIONAHA.113.01275]
54. Luciano GL, Brennan MJ, Rothberg MB. Postprandial hypotension. Am J Med. Mar 2010;123(3):281. [doi: 10.1016/j.amjmed.2009.06.026] [Medline: 20193838]
55. Ali O. Genetics of type 2 diabetes. World J Diabetes. Aug 15, 2013;4(4):114-123. [doi: 10.4239/wjd.v4.i4.114] [Medline: 23961321]
56. Li C, Yang Y, Liu X, Li Z, Liu H, Tan Q. Glucose metabolism-related gene polymorphisms as the risk predictors of type 2 diabetes. Diabetol Metab Syndr. Dec 2020;12(1):97. [doi: 10.1186/s13098-020-00604-5]

## Abbreviations

**BP:** blood pressure
**DBP:** diastolic blood pressure
**ECG:** electrocardiogram
**IDH:** isolated diastolic hypertension
**ISH:** isolated systolic hypertension
**NCD:** noncommunicable disease
**RBG:** random blood glucose
**ROC-AUC:** receiver operating characteristic–area under the curve
**SBP:** systolic blood pressure
**SMOTE:** synthetic minority oversampling technique

*Edited by Ching Nam Hang; peer-reviewed by Akhil Chaturvedi, Fakhare Alam, Tarek Abd El-Hafeez; submitted 01.02.2024; final revised version received 06.04.2024; accepted 24.04.2024; published 11.09.2024*

*Please cite as:*
*Oyebola K, Ligali F, Owoloye A, Erinwusi B, Alo Y, Musa AZ, Aina O, Salako B*
*Machine Learning–Based Hyperglycemia Prediction: Enhancing Risk Assessment in a Cohort of Undiagnosed Individuals*
*JMIRx Med 2024;5:e56993*
*URL: https://med.jmirx.org/2024/1/e56993*
*doi: 10.2196/56993*