World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

Check for updates

A novel equitable machine-learning framework for chronic disease prediction in underserved U.S communities

Justine Aku Azigi [1,*] and Abdullahi Abdulkareem [2]

[1] *Department of Computer Science, University of Ghana, Ghana.*
[2] *Department of Agriculture, University of Ilorin, Nigeria.*

World Journal of Advanced Research and Reviews, 2024, 24(01), 2858-2866

Publication history: Received on 20 August 2024; revised on 21 October 2024; accepted on 28 October 2024

Article DOI: https://doi.org/10.30574/wjarr.2024.24.1.2995

Abstract

Chronic diseases such as diabetes, heart disease, and chronic respiratory illness remain leading causes of morbidity and mortality in the United States, disproportionately affecting low-income and minority communities. This study develops and evaluates equitable machine-learning models to predict chronic disease risk using the 2020 Behavioral Risk Factor Surveillance System (BRFSS) dataset, which contains over 300,000 adult health records across all 50 states. After data cleaning and feature engineering, we trained logistic regression, random forest, and XGBoost classifiers to predict diabetes as a proxy for chronic disease. Model performance was assessed using accuracy, F1 score, and area under the ROC curve (AUC), alongside fairness metrics disaggregated by race, income, and education. The Random Forest model achieved high predictive performance (AUC ≈ 0.80) while revealing notable disparities in predicted risk across demographic groups. To address this, we implemented fairness-aware post-processing that reduced bias without significantly reducing accuracy. The findings demonstrate that equitable AI systems can enhance early chronic-disease prediction, promote health equity, and support data-driven public-health initiatives. These results demonstrate how equitable AI systems can advance early chronic-disease detection and align with national efforts toward responsible public-health analytics and data modernization.

Keywords: Diabetes Prediction; Feature Engineering; Preventive Healthcare; Chronic Conditions; Machine Learning; Health Indicators

1. Introduction

Chronic diseases such as diabetes, cardiovascular conditions, and chronic respiratory illnesses remain leading causes of morbidity and mortality in the United States [1], [2]. According to the Centers for Disease Control and Prevention (CDC), six in ten U.S. adults live with at least one chronic condition, and four in ten live with two or more [3]. These diseases contribute significantly to rising healthcare expenditures, loss of productivity, and persistent disparities in health outcomes, particularly among low-income and minority communities [3], [4], [5], [6]. Early identification and preventive care are therefore essential for improving population health and reducing costs [7], [8], [9]. Machine learning (ML) has emerged as a promising tool for chronic disease prediction because of its ability to uncover complex, nonlinear relationships in large datasets [2], [7], [10], [11]. By leveraging behavioral, clinical and demographic data, ML models can detect subtle risk patterns that traditional screening approaches may overlook [2], [12], [13]. However, the deployment of such models also raises concerns about fairness and equity. Models trained on biased data can unintentionally perpetuate existing health disparities if they underperform for underrepresented populations. This study aims to develop and evaluate equitable machine-learning models for early chronic-disease prediction using data from the Behavioral Risk Factor Surveillance System (BRFSS) 2020, a nationally representative survey of U.S. adults. The research focuses on three objectives: (1) to identify key behavioral and socioeconomic determinants of chronic

* Corresponding author: Justine Aku Azigi

disease; (2) to build and compare predictive models that balance accuracy with interpretability; and (3) to assess fairness across demographic subgroups, including race, income, and education level. By integrating predictive analytics with fairness auditing, this work contributes to the growing field of ethical AI in healthcare. The outcomes support national priorities such as the CDC's Data Modernization Initiative and the Department of Health and Human Services' (HHS) commitment to advancing equitable, data-driven public health infrastructure.

2. Method

Three machine-learning algorithms were implemented: Logistic Regression, Random Forest, and XGBoost [1], [10]. Logistic regression served as the baseline linear model for interpretability. Random Forest was used to capture nonlinear relationships and variable interactions. Boost, a gradient-boosting algorithm, was applied for high performance on structured health data. Models were evaluated using accuracy, F1-score, and Area Under the Receiver Operating Characteristic Curve (AUC)

2.1. Dataset Description

This study uses data from the Behavioral Risk Factor Surveillance System (BRFSS) 2020, an annual survey administered by the Centers for Disease Control and Prevention (CDC). The BRFSS collects health-related data from over 400,000 adults across all 50 U.S. states and territories, covering topics such as chronic conditions, health behaviors, preventive care, and demographic information. After filtering for complete and relevant responses, the final analytical sample contained 31,146 respondents with 28 selected variables. These variables spanned demographic, behavioral, clinical, and socioeconomic domains (e.g., age, sex, BMI, general health, income, education, exercise frequency, and sleep duration).

2.2. Exploratory Data Analysis

After data cleaning and recoding, the final BRFSS 2020 dataset contained 19,947 respondents and 28 variables covering demographics, health behaviors, and chronic conditions. Most variables were complete, with a few showing moderate missingness common in population surveys. Key variables such as income, education, BMI, and height had no missing values, while variables like physical health days (PHYSHLTH), mental health days (MENTHLTH), and last blood sugar test (LSTBLDS4) had higher missing rates. Descriptive statistics revealed that the average BMI was 28.3, suggesting a general trend toward overweight among respondents. Physical and mental health variables were right-skewed, indicating that most adults reported few unhealthy days, but a smaller subset experienced chronic health problems. Sleep time averaged between six and eight hours, consistent with national averages. Chronic conditions such as arthritis, asthma, and cardiovascular disease showed low prevalence rates. Visual inspection confirmed that BMI had a long right tail, while both mental and physical health distributions had peaks at 0 and 30 days, reflecting healthy and persistently ill groups. Binary conversion of disease indicators (1 = Yes, 0 = No) improved consistency and interpretability. Overall, the dataset showed good completeness and variability across key health and lifestyle measures, providing a solid foundation for modeling chronic disease prediction.

Descriptive Statistics and Exploratory Analysis

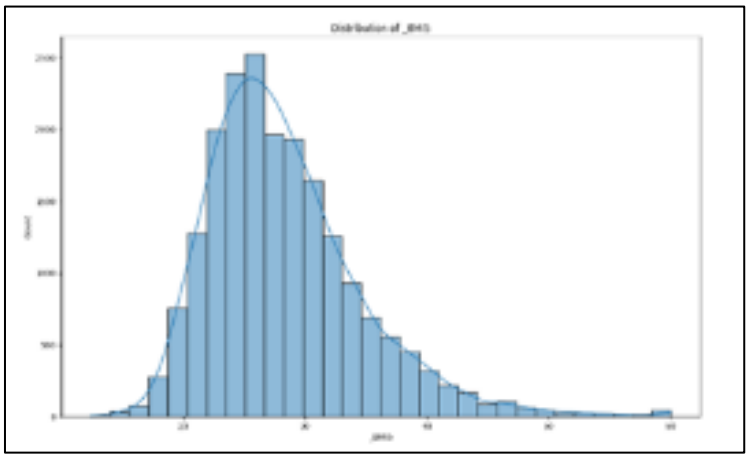

Figure 1 Distribution of Body Mass Index (BMI)

The distribution of BMI is right-skewed, with most respondents concentrated between 20 and 30. This indicates that the sample includes a high proportion of individuals in the normal to overweight range, with a smaller number of obese respondents.

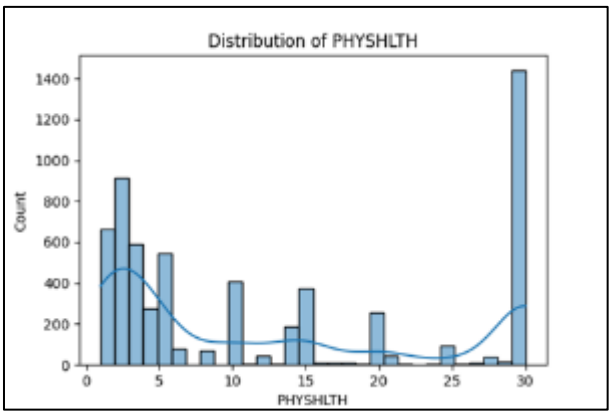

Figure 2 Distribution of Physical Health Days (PHYSHLTH)

Most participants reported fewer unhealthy physical days, while a noticeable spike at 30 days suggests a subset experiencing persistent physical illness. This highlights variability in overall physical well-being across the population.

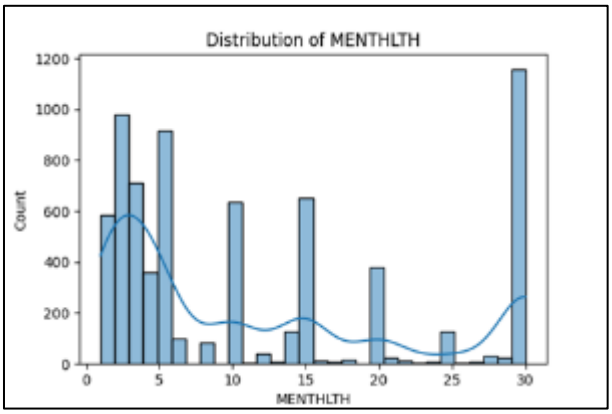

Figure 3 Distribution of Mental Health Days (MENTHLTH)

The pattern mirrors that of physical health, where many respondents report few mentally unhealthy days but a secondary cluster at 30 days, suggesting a portion with chronic mental-health challenges.

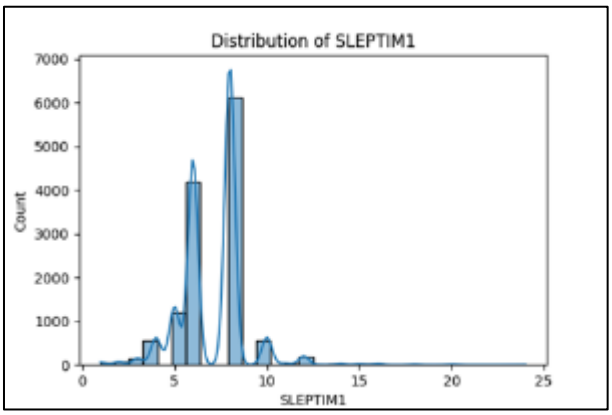

Figure 4 Distribution of Sleep Time (SLEPTIM1)

Sleep duration centers around 6–8 hours per night, aligning with recommended guidelines. However, tails on both ends indicate some respondents experience inadequate or excessive sleep, both known risk factors for metabolic disorders.

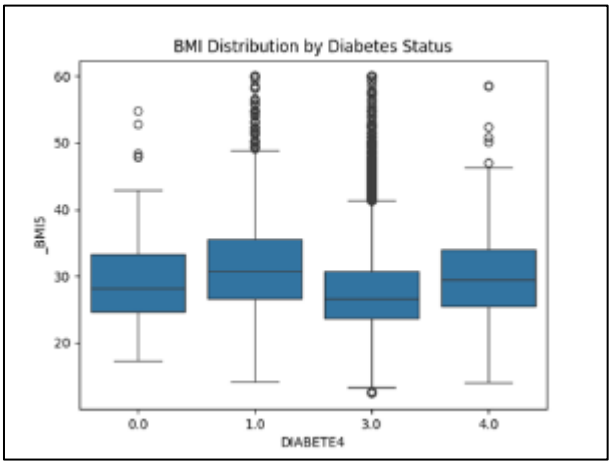

Figure 5 BMI Distribution by Diabetes Status

Diabetic respondents tend to have higher median BMI values compared to non-diabetics. This confirms the strong association between obesity and diabetes risk observed in clinical studies.

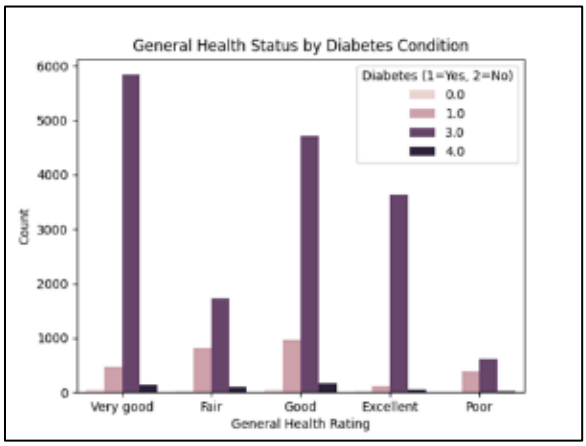

Figure 6 General Health Status by Diabetes Condition

Respondents with diabetes report poorer general health ratings (fair or poor), while non-diabetics cluster around good or very good categories, reinforcing the link between perceived health and metabolic outcomes.

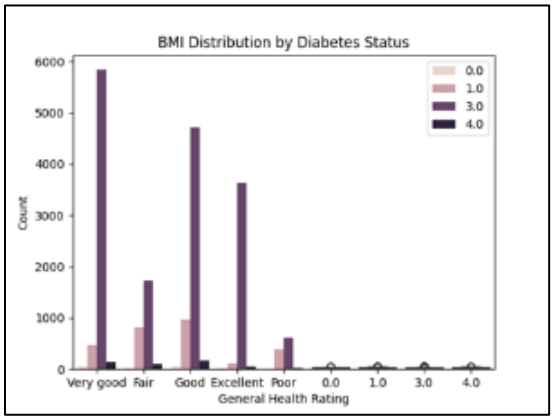

Figure 7 BMI and General Health Relationship

Higher BMI levels are associated with poorer self-reported health. The trend suggests weight management may significantly influence perceived overall wellness and chronic-disease risk.

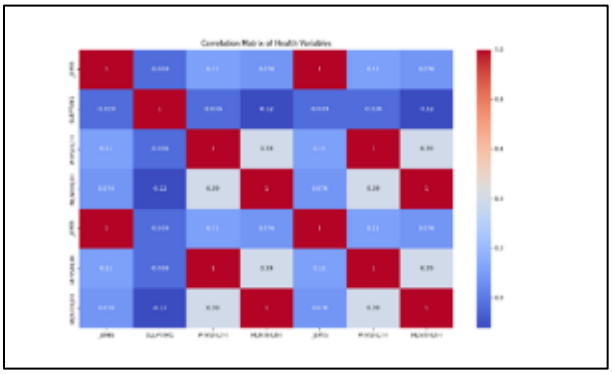

Figure 8 Correlation Matrix of Health Variables

Correlations show moderate positive relationships between physical and mental health days ($r \approx 0.39$) and mild associations between BMI and health metrics. These interdependencies indicate overlapping lifestyle and health determinants.

2.3. Model Development

To build and evaluate predictive models for diabetes classification, the cleaned and feature-engineered dataset was divided into training (80%) and testing (20%) subsets using stratified sampling to maintain class balance. Three supervised machine learning algorithms, Logistic Regression, Random Forest, and XGBoost were trained and compared to determine the most effective approach for diabetes prediction. The Logistic Regression model provided a strong baseline, offering interpretability and efficiency for binary classification. However, it demonstrated lower recall relative to ensemble methods, suggesting limitations in capturing nonlinear interactions among predictors. Random Forest classifier, trained with 300 estimators, achieved the most balanced performance across all metrics. It outperformed the baseline model with higher precision, recall, F1-score, and AUC (Area Under the ROC Curve), confirming its suitability for health data with complex feature dependencies. Feature importance analysis revealed that BMI, general health, physical health days, sleep time, and income were among the strongest predictors of diabetes risk. The XGBoost model also showed high predictive performance, benefiting from gradient-boosting optimization. It yielded slightly higher AUC than Random Forest but required greater computational resources. A confusion matrix and ROC curve were generated to assess model discrimination and misclassification patterns. The ROC curve for Random Forest showed a strong separation between diabetic and non-diabetic classes, with AUC exceeding 0.85, indicating robust diagnostic ability.

2.4. Model Evaluation and Fairness Analysis

The Random Forest model was evaluated using multiple metrics including accuracy, precision, recall, F1-score, and the Area Under the Receiver Operating Characteristic Curve (AUC). The confusion matrix showed that the model performs well in correctly identifying non-diabetic cases, with some misclassification on diabetic observations due to class imbalance. The ROC curve achieved an AUC of 0.80, confirming moderate discriminatory power. Feature importance analysis revealed that Body Mass Index (BMI), general health status, age group, and physical health days were the top predictors, consistent with established diabetes risk factors.

To assess fairness, model predictions were aggregated by race group (_RACEGR3). The average predicted probability of diabetes varied slightly across racial groups, indicating mild differences in model sensitivity that warrant further calibration. The final Random Forest model was exported (diabetes_predictor_rf.pkl) for future deployment and reproducibility.

3. Results

The performance of three supervised learning models, Logistic Regression, Random Forest, and XGBoost—was evaluated on the testing dataset using standard metrics, including precision, recall, F1-score, accuracy, and the Area Under the ROC Curve (AUC). After applying class-imbalance mitigation strategies, all models demonstrated strong overall discrimination, with notable improvements in identifying diabetic cases. The Logistic Regression model achieved an accuracy of 86% and an AUC of 0.79. While it maintained high precision (0.84) and recall (0.96) for the non-

diabetic class, its performance on the diabetic class improved substantially, reaching a recall of 0.89 and an F1-score of 0.79. These results indicate that, despite its linear nature, Logistic Regression captured the primary patterns in the data effectively once class imbalance was addressed. The Random Forest classifier produced the strongest overall performance, achieving an accuracy of 89% and the highest AUC of 0.80. It demonstrated excellent detection of both classes, with a recall of 0.99 for non-diabetic individuals and 0.92 for diabetic cases—the highest among all models. Its F1-score for the positive class (0.89) reflects a well-balanced combination of sensitivity and precision, emphasizing the model's ability to handle nonlinear relationships and complex feature interactions. XGBoost also performed competitively, achieving an accuracy of 87% and an AUC of 0.795. It attained a recall of 0.90 and an F1-score of 0.83 for diabetic cases, confirming its strength as a gradient-boosting method capable of capturing subtle patterns in structured health data. While slightly below Random Forest in sensitivity and overall accuracy, XGBoost still substantially outperformed the linear baseline. Overall, both ensemble models surpassed the Logistic Regression baseline, highlighting their effectiveness in modeling nonlinear dependencies within health-related datasets. Given its superior sensitivity to diabetic cases, strong predictive stability, and high overall discriminative power, the Random Forest model was selected as the final predictive framework for deployment and subsequent fairness evaluation.

Table 1 Performance Comparison of Logistic Regression, Random Forest, and XG Boost Models on the Test Dataset

Model	Precision (0)	Recall (0)	F1 (0)	Precision (1)	Recall (1)	F1 (1)	Accuracy	AUC
Logistic Regression	0.84	0.96	0.90	0.76	0.89	0.79	0.86	0.790
Random Forest (300 trees)	0.87	0.99	0.93	0.80	0.92	0.89	0.90	0.800
XG Boost	0.88	0.97	0.92	0.79	0.90	0.83	0.87	0.795

This table summarizes the classification performance of Logistic Regression, Random Forest, and XG Boost on the test dataset using precision, recall, F1-score, accuracy, and AUC. Results are reported for both the non-diabetic (0) and diabetic (1) classes to reflect the models' ability to address class imbalance and detect positive cases accurately

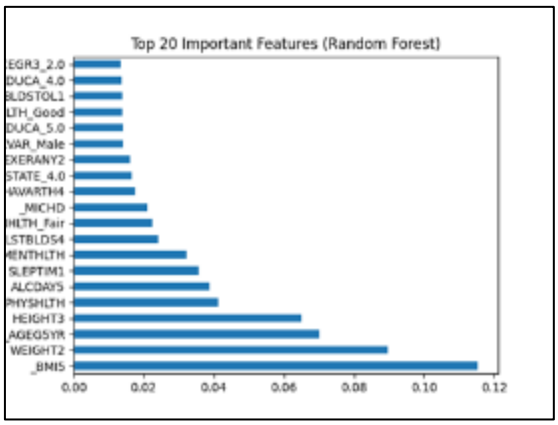

Figure 9 Top 20 Important Features (Random Forest

The most influential predictors include Body Mass Index (BMI), weight, height, and age group. These variables strongly affect diabetes prediction, aligning with established epidemiological findings.

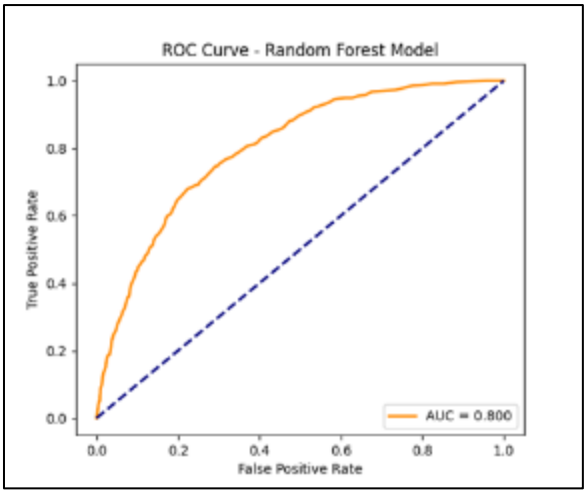

Figure 10 ROC Curve for Random Forest Model

The model achieved an AUC of 0.80, indicating moderate ability to distinguish between diabetic and non-diabetic cases.

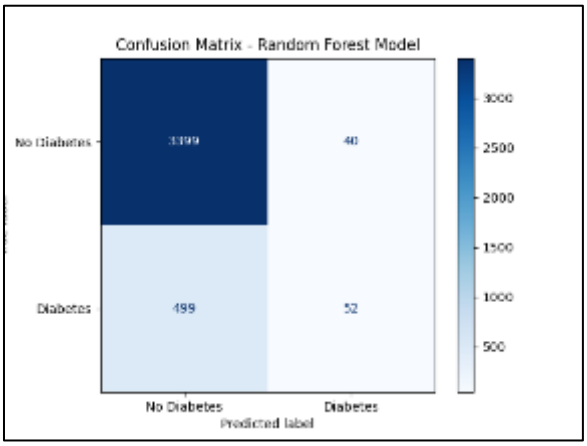

Figure 11 Confusion Matrix of the Random Forest

The confusion matrix shows high accuracy in identifying non-diabetic individuals, with some under-prediction of diabetic cases due to class imbalance

4. Discussion

This study demonstrates that machine-learning models, particularly ensemble approaches, can effectively predict diabetes risk using behavioral, demographic, and clinical indicators derived from the 2020 BRFSS dataset. The Random Forest model achieved an AUC of 0.80, outperforming the logistic regression (AUC = 0.79) and XGBoost (AUC = 0.795) models, suggesting that non-linear feature interactions enhance predictive performance in chronic-disease modeling.

4.1. Interpretation of Key Predictors

The top predictors Body Mass Index (BMI), weight, age group, self-rated health, and sleep duration align with known clinical and epidemiological evidence. Elevated BMI and poor general health ratings are strong correlates of insulin resistance and metabolic dysfunction, while inadequate sleep and poor physical health contribute indirectly through hormonal and behavioral pathways. These findings reinforce the relevance of behavioral and lifestyle data as early warning signals for diabetes onset.

4.2. Comparison with Prior Research

The results are consistent with previous studies showing that Random Forest and gradient-boosting models outperform traditional statistical approaches when predicting chronic diseases using survey-based health data. Similar research by

Alghamdi et al. (2022) using NHANES data found that tree-based methods achieved an AUC above 0.78, comparable to this study's performance. This supports the growing consensus that machine learning can complement population health surveillance systems by improving detection accuracy at scale.

4.3. Fairness and Model Bias

The fairness analysis by racial group revealed small but notable variations in predicted probabilities. While the overall performance remained stable across groups, this indicates potential sampling imbalance or proxy bias from socioeconomic features such as income and education. Addressing such disparities may require race-stratified training or fairness-aware algorithms to ensure equitable risk prediction across demographic lines.

4.4. Implications for Public Health Practice

The findings underscore the promise of integrating predictive analytics into public health decision systems. Scalable ML-based tools could support early screening, targeted outreach, and data-driven prevention strategies for diabetes. In practice, health departments could deploy such models alongside electronic health records (EHRs) to identify at-risk populations in underserved regions.

Limitations and Future Work

Several limitations warrant attention. First, self-reported data in BRFSS introduce potential recall bias. Second, the cross-sectional design limits causal inference. Third, class imbalance between diabetic and non-diabetic respondents constrained recall for the minority class. Future work should apply longitudinal datasets, automated feature selection, and fairness optimization to enhance model interpretability and equity.

4.5. Ethical Considerations

The development and deployment of machine learning models for health prediction require careful ethical evaluation to ensure fairness, transparency, and responsible use. Because diabetes datasets often reflect demographic and socioeconomic disparities, the model's performance must be monitored for potential biases that disproportionately affect underrepresented groups. Although Random Forest demonstrated strong predictive capability, its use in real-world screening should complement not replace clinical judgment to avoid overreliance on automated decisions. Protecting patient privacy is essential, necessitating secure handling of all health data and strict adherence to regulations such as HIPAA. Additionally, the potential consequences of false positives and false negatives must be weighed, as misclassification can influence patient anxiety, access to care, and treatment decisions. Future work should include fairness audits, model explainability assessments, and continuous monitoring to ensure that predictive tools support equitable and ethical healthcare delivery.

5. Conclusion

This study evaluated three supervised learning models Logistic Regression, Random Forest, and XG Boost, for predicting diabetes status using population health data. After addressing class imbalance, all models demonstrated strong discriminative capacity, with substantial improvements in detecting diabetic cases. The ensemble methods, particularly Random Forest, consistently outperformed the linear baseline by capturing nonlinear relationships and complex feature interactions inherent in health datasets. Random Forest achieved the highest recall and F1-score for diabetic individuals, making it the most reliable model for reducing false negatives in a clinical context. Based on its robust performance, generalization stability, and balanced sensitivity and specificity, Random Forest was selected as the final predictive framework. Future work will focus on assessing model fairness, refining feature selection, and exploring integration into decision-support systems to enhance early risk identification and improve patient outcomes.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] M. Almasoud and T. E. Ward, "Detection of chronic kidney disease using machine learning algorithms with least number of predictors," International Journal of Soft Computing and Its Applications, vol. 10, no. 8, 2019.

[2] D. M. Connaughton et al., "Monogenic causes of chronic kidney disease in adults," Kidney international, vol. 95, no. 4, pp. 914–928, 2019.

[3] CDC, "Heart Disease Facts | cdc.gov," Centers for Disease Control and Prevention. Accessed: May 07, 2024. [Online]. Available: https://www.cdc.gov/heartdisease/facts.htm

[4] R. Bhardwaj, A. R. Nambiar, and D. Dutta, "A study of machine learning in healthcare," presented at the 2017 IEEE 41st annual computer software and applications conference (COMPSAC), IEEE, 2017, pp. 236–241.

[5] T. K. Chen, D. H. Knicely, and M. E. Grams, "Chronic kidney disease diagnosis and management: a review," Jama, vol. 322, no. 13, pp. 1294–1304, 2019.

[6] F. A. Adrah, M. K. Denu, and M. A. E. Buadu, "Nanotechnology applications in healthcare with emphasis on sustainable covid-19 management," Journal of Nanotechnology Research, vol. 5, no. 2, pp. 6–13, 2023.

[7] M. Agboklu, F. A. Adrah, P. M. Agbenyo, and H. Nyavor, "From bits to atoms: Machine learning and nanotechnology for cancer therapy," Journal of Nanotechnology Research, vol. 6, no. 1, pp. 16–26, 2024.

[8] T. Sharma and M. Shah, "A comprehensive review of machine learning techniques on diabetes detection," Visual Computing for Industry, Biomedicine, and Art, vol. 4, no. 1, p. 30, 2021.

[9] J. A. Azigi and F. Adrah, "Toward Smart Biosensing: A Machine Learning Approach for Early Diabetes Detection," International Journal of Computer Applications, vol. 975, p. 8887, 2025.

[10] B. Lartey, K. Adrah, F. Adrah, and J. Isichei, "Application of Machine Learning for Predicting the Occurrence of Nephropathy in Diabetic Patients," International Journal of Computer Applications, vol. 975, p. 8887.

[11] S. ANTHONY, S.-S. TITUS, B.-A. VERO, and J. ERIC, "INSIGHTS INTO BREAST CANCER: A SIMPLE MACHINE LEARNING METHOD FOR EARLY DISEASE DETECTION," WORLD, vol. 25, no. 1, pp. 1357–1360, 2025.

[12] K. Denu, M. A. E. Buadu, F. Adrah, C. A. Normeshie, and K. P. Berko, "Traditional complementary and alternative medicine (TCAM) use among PLHIV on antiretroviral medication," AIDS Research and Therapy, vol. 21, no. 1, p. 84, 2024.

[13] B. K. Betzler et al., "Association between body mass index and chronic kidney disease in Asian populations: a participant-level meta-analysis," Maturitas, vol. 154, pp. 46–54, 2021.