A Decision Tree Ensemble Approach to Diabetes Prediction using the Framingham Heart Dataset, Exploring the Role of AI-Associated Interventions in Reducing Diabetes-Related Adverse Outcomes Between Men and Women

Patricia Y. Talbert[1,*], Korin Reid[2], Donald Smith[3]

[1]College of Nursing and Allied Health Sciences, Howard University

[2]Ellison Laboratory

[3]AI Discovery Center of Alabama

Research Article

Open Access &

Peer-Reviewed Article

DOI: 10.14302/issn.2641-4538.jphi-25-5886

Corresponding author:

Patricia Y. Talbert, College of Nursing and Allied Health Sciences, Howard University

Keywords:

Diabetes risk prediction, AI-driven interventions, Random Forests, Synthetic Minority Over-sampling Technique (SMOTE)

Received: November 28, 2025

Accepted: December 15, 2025

Published: December 29, 2025

Academic Editor:

Sasho Stoleski, Institute of Occupational Health of R. Macedonia, WHO CC and Ga2len CC

Citation:

Patricia Y. Talbert, Korin Reid, Donald Smith (2025) A Decision Tree Ensemble Approach to Diabetes Prediction using the Framingham Heart Dataset, Exploring the Role of AI-Associated Interventions in Reducing Diabetes-Related Adverse Outcomes Between Men and Women. Journal of Public Health International - 7(4):28-33. https://doi.org/10.14302/issn.2641-4538.jphi-25-5886

Abstract

Objective

Diabetes poses significant public health challenges, with many individuals remaining undiagnosed and at risk of complications. This study aimed to evaluate the performance of decision tree ensemble methods for predicting diabetes onset using the Framingham Heart Study Teaching Dataset and to explore sex-specific risk patterns relevant to AI-driven interventions.

Methods

We analyzed data from 11,627 participants, incorporating demographics, vital signs, smoking status, medication use, and laboratory measures. Random Forest classifiers were developed to predict diabetes incidence at approximately 6-year (Period 2) and 12-year (Period 3) follow-ups. Class imbalance was addressed using undersampling, oversampling, and the Synthetic Minority Over-sampling Technique (SMOTE).

Results

The models demonstrated robust performance, achieving an Area Under the Curve (AUC) of 0.856 in Period 2, and moderate predictive ability in Period 3 (AUC = 0.732 in males, 0.786 in females). Key predictors included glucose level, BMI, systolic blood pressure, age, and heart rate. Notably, differences emerged in predictive accuracy between men and women, suggesting potential sex-specific vulnerabilities that merit further study.

Conclusion

Machine learning approaches, particularly Random Forests, show promise for medium- and long-term diabetes risk prediction, supporting early iden-

tification and intervention efforts. Future work should focus on hyperparameter tuning and explainability techniques, such as SHapley Additive exPlanations (SHAP) values, to improve model precision, interpretability, and fairness. Equity-focused strategies remain critical to ensure AI-driven tools benefit diverse populations and do not exacerbate existing disparities in diabetes care.

Introduction

Diabetes affects approximately 11.6 percent of the United States population, representing a substantial public health burden with wide-ranging implications for individuals and healthcare systems alike [1]. Alarmingly, an estimated 22.8 percent of adults living with diabetes are unaware of their condition, which delays treatment, increases the risk of serious complications, and contributes to rising healthcare costs [1]. For instance, large datasets from electronic health records (EHR) and other sources can be leveraged to identify high-risk individuals, prompting earlier clinical follow-up. To that end, this study utilizes the Framingham Heart Study for diabetes prediction using machine learning-based methods. Numerous prior works use the dataset for diabetes related studies. Ding et al. (2024) demonstrated that elevated blood glucose and insulin levels were more strongly associated with cognitive decline in women, pointing to sex-based vulnerability in neurocognitive outcomes.

Kanaya and colleagues (2022) extended the diabetes risk landscape to include increased susceptibility to cardiac arrhythmias, further emphasizing the systemic burden of the disease. Kaplan et al. (2022) found that while male sex initially appeared to predict diabetes incidence, the association diminished after adjusting for metabolic and behavioral risk factors, underscoring the influence of modifiable conditions. These studies emphasize the need for diabetes interventions integrating vascular, cognitive, and systemic risk stratification. Most similar to our research, Ai et al. (2025) use the Framingham offspring study to predict diabetes using logistic regression.

Objective

Using the Framingham Heart Study dataset, this study investigates the effectiveness and equity of artificial intelligence (AI) and machine learning (ML) interventions in predicting and reducing diabetes-related adverse outcomes among women and men. Gradient boosted decision trees were employed to predict diabetes onset. This exploration is crucial for developing personalized and equitable healthcare strategies that can significantly improve patient outcomes and reduce the burden of diabetes. Understanding how AI/ML models perform across different demographic groups can ensure that these advanced technologies benefit all individuals, fostering a more inclusive and practical approach to diabetes prevention and management. This work is even more critical as it directly addresses the growing public health crisis posed by diabetes, striving for a future with more effective prevention and management strategies for everyone.

Conceptual Framework

This study investigates the effectiveness and equity of AI/ML interventions in predicting and reducing diabetes-related adverse outcomes among women and men. Using the Framingham Heart Study dataset, gradient boosted decision trees were employed to predict diabetes onset. This exploration is crucial for developing personalized and equitable healthcare strategies that can significantly improve patient outcomes and reduce the burden of diabetes. Understanding how AI/ML models perform across different demographic groups can ensure that these advanced technologies benefit all individuals, fostering a more inclusive and practical approach to diabetes prevention and management. This work is even more critical as it directly addresses the growing public health crisis posed by diabetes, striving

for a future with more effective prevention and management strategies for everyone. AI can significantly reduce the burden of diabetes by accurately predicting its onset using advanced machine learning models, enabling early detection and timely interventions. By leveraging large datasets, AI identifies high-risk individuals, prompting personalized preventative care and lifestyle adjustments. This proactive approach helps manage the condition before it progresses, ultimately leading to better health outcomes, reduced healthcare costs, and a healthier society.

For instance, Ai et al. (2025), in their work, *Diabetes Mellitus Risk Prediction in the Framingham Offspring Study and Large Population Analysis*, published in Nutrients, demonstrate how logistic regression can be used with the Framingham offspring study to predict diabetes, highlighting AI's role in risk prediction. Another relevant and current work is by Kaplan and colleagues (2022). Their article, "Predictors of Incident Diabetes in Two Populations: Framingham Heart Study and Hispanic Community Health Study / Study of Latinos," was published in BMC Public Health in 2022. This research also contributes to understanding and predicting diabetes onset, further supporting the role of AI and advanced analytical methods in addressing diabetes mellitus. Many others are currently exploring this significant topic and showing how AI/ML can be used to change the status quo and address this issue.

Methods

We applied decision tree ensemble methods to predict diabetes onset using the Framingham Heart Study Teaching Dataset (N=11,627), accessed via Terra's cloud platform. Predictive features included demographics, vital signs, smoking status, medication use, and lab measurements. Table 1 provides an overview of the variables explored for this study. The Framingham Heart dataset includes baseline measurements and two follow-up periods, on average, 6 years apart. We use a random forest to predict diabetes incidence in each follow-up period. To address class imbalance, we use undersampling, oversampling, and SMOTE. Our Random Forest models, adjusted for class imbalance via undersampling, oversampling, and the synthetic minority oversampling technique (SMOTE), achieved Area Under the Curve (AUC) > 0.8 for ~6-year predictions (Period 2) and >0.76 for ~12-year predictions (Period 3). Future work will involve hyperparameter tuning and SHapley Additive exPlanations (SHAP)-based interpretability to enhance model precision and fairness. Our Random Forest models demonstrated promising predictive performance for diabetes onset across two follow-up periods. In **Period 2** (approximately 2–4 years post-baseline), the model achieved an **AUC of 0.856**, indicating strong discriminative ability. Precision was **0.158**, and recall was **0.688**. Future work will examine whether we can optimize parameters to reduce false positives while maintaining reasonable recall. For **Period 3**, representing a longer-term prediction window, model performance was more modest, with an overall AUC of **0.732** for males and **0.786** for females. Notably, **glucose level**, **BMI**, **systolic blood pressure**, **age**, and **heart rate** emerged as the top predictors, as illustrated in Table 2. These results underscore the potential of machine learning models to support early diabetes risk identification while highlighting the need for further refinement to enhance precision and minimize false negatives.

Discussion

These findings revealed that Table 1 summarizes the variables from the Framingham Heart Study dataset, including demographic, clinical, and lifestyle factors incorporated into the model. Table 2 summarizes model evaluation results across two prediction periods using baseline data from the Framingham Heart Study. Metrics include precision, recall, and area under the Receiver Operating Characteristic **(**ROC) curve (AUC), reported separately for male, female, and overall populations. Results demonstrate consistent model performance with slight variations by sex and time horizon. With machine

Table 1. List of Predictor Variables and Descriptions Used in the Diabetes Prediction Models

Study Variable	Description
Age	Age of participants
Sex	Biological sex (1 = male, 0 = female)
BMI	Body Mass Index
SYSBP	Systolic blood pressure
DIABP	Diastolic blood pressure
CURSMOKE	Current smoking status (1 = yes, 0 = no)
CIGPDAY	Cigarettes smoked per day
BPMEDS	On blood pressure medication (1 = yes, 0 = no)
HEARTRTE	Heart rate
GLUCOSE	Glucose level
PREVCHD	Previous coronary heart disease (1 = yes, 0 = no)
PREVHYP	Previous hypertension (1 = yes, 0 = no)
HDLC	HDL cholesterol
LDLC	LDL cholesterol

Table 2. Model Performance by Group and Follow-up Period using Baseline Values

Period	Group	Patients	Diabetes Cases (%)	Precision	Recall	AUC
Period 2	All	795	32 (4.0%)	0.158	0.688	0.856
Period 2	Male	399	18 (4.5%)	0.169	0.667	0.831
Period 2	Female	488	20 (4.1%)	0.146	0.65	0.835
Period 3	Male	399	31 (7.8%)	0.147	0.645	0.732
Period 3	Female	488	35 (7.2%)	0.192	0.686	0.786

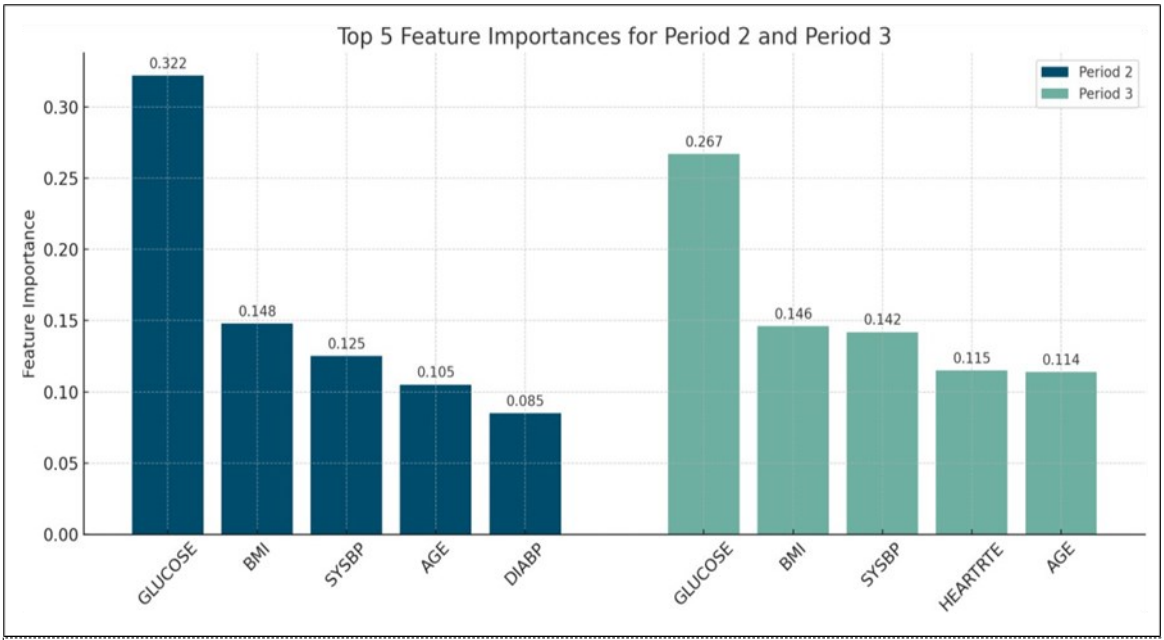

Figure 1. Top Five Feature Importances for Diabetes Prediction Models at Periods 2 and 3

learning, the ROC curve is a graphical tool used to evaluate the performance of a binary classification model—a model that predicts one of two possible outcomes, such as the presence or absence of a disease. The ROC curve helps us assess how effectively the model distinguishes between *positive cases* (for example, individuals who have the disease) and *negative cases* (those who do not) across various *threshold levels*.

Each threshold represents a decision point where the model decides whether to classify an observation as positive or negative based on the predicted probability [6, 7, 8]. Likewise, by plotting the *True Positive Rate (Sensitivity*) on the y- axis against the *False Positive Rate (1 - Specificity)* on the x-axis for all threshold values, the ROC curve provides a visual representation of the model's ability to differentiate between the two classes. The closer the curve follows the top-left border of the graph, the better the model is at distinguishing between positive and negative outcomes. This visualization ultimately allows us to compare different models and select the one with the best overall discriminatory power [6, 7, 8].

In the findings, Figure 1 provides the top 5 important features for predicting diabetes onset in Periods 2 and 3. Glucose level consistently emerged as the strongest predictor across both periods, followed by BMI, systolic blood pressure (SYSBP), age, and either diastolic blood pressure (DIABP) or heart rate (HEARTRTE). The chart illustrates distinct contributions of these features to model performance in each follow-up period.

Conclusion

Our study demonstrates that machine learning models, specifically Random Forest classifiers, can effectively predict the onset of diabetes using data from the Framingham Heart Study. The models performed well in the medium-term prediction window (Period 2), with an AUC of 0.856, and showed moderate predictive ability for longer-term risk (Period 3). Key predictors consistently included glucose levels, BMI, systolic blood pressure, age, and heart rate, which align with established diabetes risk factors. The results highlight the importance of addressing class imbalance—our use of random undersampling contributed to more balanced sensitivity and specificity, particularly in controlling false nega-

tives, which is critical for timely intervention. While these findings are promising, there is an opportunity to further enhance model precision through hyperparameter tuning and incorporating advanced explainability techniques such as SHAP values. These future steps will help improve model transparency, interpretability, and fairness, supporting the development of equitable, data-driven strategies for early diabetes risk detection in diverse populations. AI/ML-enabled diabetes interventions show promise in improving inpatient outcomes but may not be equitably accessed or equally beneficial across all populations. Equity-focused design, implementation, and evaluation of AI tools are essential to ensure that technological advancements do not exacerbate existing disparities in diabetes care.

Acknowledgment

The research reported in this manuscript was supported by Howard University, AI Discovery Center of Alabama, Inc., and AIM-AHEAD Coordinating Center, award number OTA-21-017, and was, in part, funded by the National Institutes of Health Agreement No. 1OT2OD032581.

The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the NIH.

References

1. Centers for Disease Control and Prevention. ([CDC], 2024). National Diabetes Statistics Report. Diabetes, https://www.cdc.gov/diabetes/php/data- research/index.html#:~:text=8.7%20million%20adults%20aged%2018,all%20U.S.%20ad ults%20with%20diabetes.
2. Ding, H., Liu, C., Li, Y., Fang, T., Ang, A., Devine, S., et al (2024). Sex-Specific Blood Biomarkers Linked to Memory Changes in Middle-Aged Adults: The Framingham Heart Study. *Alzheimer's & Dementia (Amsterdam, Netherlands) 16*(1): e12569.
3. Kanaya, A. M, Grady D., & Barrett-Connor, E. (2002). Explaining the sex difference in coronary heart disease mortality among patients with type 2 diabetes mellitus: a meta- analysis. *Archives Internal Medicine*, *162*(15):1737–1745.
4. Kaplan, R. C., Song, R. J., Lin, J., Xanthakis, V., Hua, S., Chernofsky, A. *et al.* (2022). Predictors of incident diabetes in two populations: Framingham Heart Study and Hispanic Community Health Study/Study of Latinos. *BMC Public Health 22*, 1053. https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-022-13463-8
5. Ai, M., Otokozawa, S., Liu, C.T., Asztalos, B. F, Maddalena, J., Diffenderfer, M. R., et al. (2025, March) Diabetes Mellitus Risk Prediction in the Framingham Offspring Study and Large Population Analysis. *Nutrients. 24*;17(7):1117. doi: 10.3390/nu17071117. PMID: 40218874; PMCID: PMC11990307.
6. Li, J. (2024). Area under the ROC curve has the most consistent evaluation for binary classification. *PLOS ONE*, *19*(12), e0316019. https://doi.org/10.1371/journal.pone.0316019
7. Muschelli, J. (2019). ROC and AUC with a binary predictor: A potentially misleading metric. *Journal of Classification*, 37(3), 696-708. https://doi.org/10.1007/s00357-019-09345-1
8. Yin, J., & Vogel, R. L. (2017). Using the ROC curve to measure association and evaluate prediction accuracy for a binary outcome. *Biometrics & Biostatistics International Journal*, 5(3), 1-10. https://doi.org/10.15406/bbij.2017.05.00134