Proceedings of the 4th International Conference on Applied Physics and Mathematical Modeling
DOI: 10.54254/2753-8818/2026.35377

Non-Laboratory Variables for Adult Diabetes Risk Screening: A Comparison of Random Forest and XGBoost Using NHANES 2017–March 2020

Siqi Jin

School of Statistics, East China Normal University, Shanghai, China
10230750419@stu.ecnu.edu.cn

Abstract. Diabetes is a growing chronic disease burden, and early screening is important for identifying individuals who may require further clinical testing. However, laboratory biomarkers such as glycated hemoglobin (HbA1c) may not always be available in low-cost population screening. Using the NHANES 2017-March 2020 pre-pandemic public-use data, we examined whether non-laboratory variables can support diabetes risk screening in adults. Adults aged 20 years or older were included. Diabetes status was defined using self-reported doctor diagnosis and HbA1c-based outcome labelling, while laboratory biomarkers were excluded from the main predictor set. The main non-laboratory predictors included demographic, socioeconomic, anthropometric, blood pressure, and lifestyle variables. Two tree-based machine learning models, Random Forest and XGBoost, were trained and evaluated on a testing set. Model performance was assessed using AUC, accuracy, sensitivity, specificity, F1-score, and Brier score. The final main sample contained 5,728 adults, including 1,071 diabetes cases. In the non-laboratory setting, XGBoost achieved higher AUC and much higher sensitivity than Random Forest. Adding HbA1c in a supplementary model substantially improved classification performance. These findings suggest that non-laboratory variables may support preliminary risk screening, but laboratory testing remains important for accurate diabetes classification.

Keywords: diabetes screening, non-laboratory variables, Random Forest, XGBoost, NHANES

1. Introduction

Diabetes is an increasingly important public health issue. The International Diabetes Federation estimates that approximately 589 million adults aged 20 to 79 were living with diabetes in 2024, and this number is projected to reach 853 million by 2050 [1]. In the United States, the National Center for Health Statistics reported that the prevalence of total diabetes among adults was 15.8% during August 2021–August 2023, while the age-adjusted prevalence of total diabetes increased from 9.7% in 1999–2000 to 14.3% in August 2021–August 2023 [2]. These trends make early risk screening a relevant statistical and public health problem.

DOI: 10.54254/2753-8818/2026.35377

Clinical diagnosis of diabetes relies on information such as self-reported diagnosis, HbA1c, fasting plasma glucose, and other laboratory-based measurements. These biomarkers are informative, but they require laboratory testing and may not always be available before an individual enters the health system. For preliminary population screening, it is therefore useful to ask whether easily collected non-laboratory variables, including age, body size, blood pressure, and lifestyle indicators, can provide meaningful risk information. This approach does not aim to replace clinical diagnosis; rather, it seeks to evaluate whether low-cost variables can effectively identify individuals who would benefit from further testing.

Existing prediction studies have used machine learning methods for diabetes and related chronic diseases [3]. However, models that include HbA1c or fasting glucose as ordinary predictors may overstate screening performance, because these biomarkers are already part of the diabetes diagnostic process. This study therefore separates two settings. The main analysis uses only non-laboratory predictors to represent a low-cost screening situation. A supplementary laboratory-enhanced analysis then adds HbA1c to measure the additional value of laboratory information. We compare Random Forest and XGBoost because both are suitable for structured tabular data and can capture nonlinear relationships and interactions among predictors. The research questions are: (1) how well can non-laboratory variables classify adult diabetes status; (2) which of the two machine learning models performs better for screening; and (3) how much does performance improve when HbA1c is added?The significance of this study is that it evaluates the practical screening value of easily collected variables while clarifying the additional contribution of laboratory biomarkers.

2. Methods

2.1. Data source and study population

We used the NHANES 2017-March 2020 pre-pandemic public-use data. NHANES combines interviews, physical examinations, and laboratory measurements in a nationally representative U.S. survey design [4]. The 2017-March 2020 pre-pandemic file was created by combining the 2017-2018 cycle with the partial 2019-March 2020 data, with analytic guidance provided by the National Center for Health Statistics [5]. Although NHANES has a complex survey design, the present analysis used a cleaned complete-case sample for predictive modelling rather than estimating nationally representative prevalence.

The study population was restricted to adults aged 20 years or older. This restriction was implemented because adult diabetes risk screening is the target setting, and variables such as BMI, waist circumference, blood pressure, smoking, and physical activity are more consistently interpreted in adults than in children and adolescents. NHANES does not directly distinguish type 1 and type 2 diabetes in the cleaned outcome used here. Therefore, the outcome should be interpreted as adult diabetes status rather than confirmed type 2 diabetes.

2.2. Outcome and predictors

The outcome was current diabetes status. Participants were classified as having diabetes if they reported a doctor's diagnosis of diabetes or met the HbA1c threshold used for outcome labelling. The HbA1c threshold was based on standard diagnostic criteria, where A1C values of 6.5% or higher are consistent with diabetes diagnosis when confirmed clinically [6]. In the main screening models, however, HbA1c was excluded from the predictor set to avoid information leakage. This

DOI: 10.54254/2753-8818/2026.35377

distinction is important: laboratory biomarkers were used to define the outcome, but the main model tested whether non-laboratory variables alone could support screening.

The non-laboratory predictor set included age, sex, race/ethnicity, education, income-to-poverty ratio, BMI, waist circumference, systolic blood pressure, diastolic blood pressure, smoking status, physical activity, and sedentary time. These variables can be obtained through questionnaires or basic physical examinations. Categorical variables such as race/ethnicity and education were converted into one-hot encoded variables for model training.

2.3. Model training and evaluation

We trained two machine learning models: Random Forest and XGBoost. Random Forest builds an ensemble of decision trees and averages their predictions, reducing the instability of a single tree while capturing nonlinear patterns [7]. XGBoost is a gradient boosting method that sequentially improves prediction by adding trees that correct previous errors, and it has been widely used for structured tabular prediction tasks [8]. These two models were selected to keep the analysis focused while still comparing two common ensemble tree methods.

The cleaned dataset was divided into training and testing subsets. Model tuning was conducted within the training process, and final performance was evaluated on the testing set. Because diabetes cases were less frequent than non-diabetes cases, we did not rely on accuracy alone. The evaluation metrics included AUC, accuracy, sensitivity, specificity, F1-score, and Brier score. Sensitivity was of particular importance, as screening programs prioritize the detection of individuals requiring follow-up tests. We also examined XGBoost variable importance to identify the non-laboratory predictors most influential for prediction.

Figure 1 presents an overview of the data processing and model comparison strategy. The main models used only non-laboratory variables, while the supplementary models added HbA1c to evaluate the additional value of laboratory information.

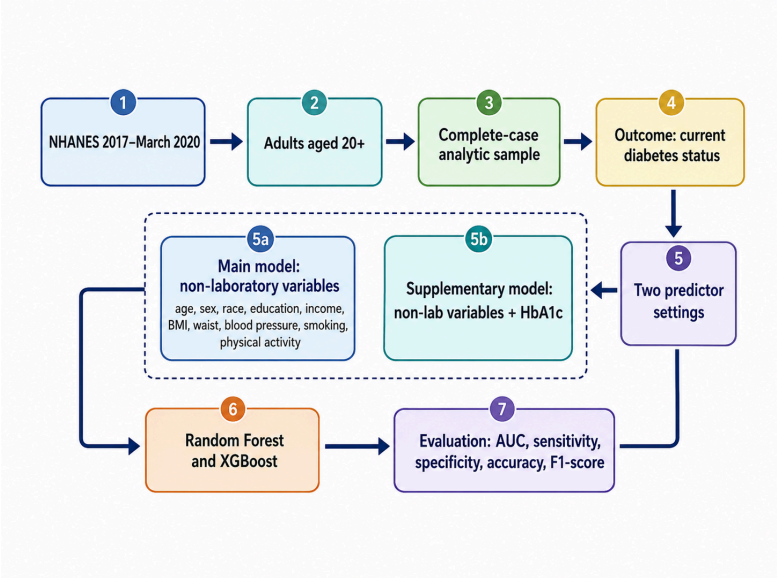

Figure 1. Study workflow and model design

DOI: 10.54254/2753-8818/2026.35377

3. Results

3.1. Analytic samples

After complete-case filtering, the main non-laboratory sample included 5,728 adults, of whom 1,071 were classified as having diabetes. Diabetes cases therefore represented 18.7% of the cleaned main sample. The laboratory-enhanced sample included 5,696 adults, including 1,039 diabetes cases, corresponding to 18.2% of that sample. The small difference between the two samples indicates that adding HbA1c caused limited additional sample loss.

Table 1 summarizes the two analytic samples. It is important to note that the diabetes proportion reported here should not be interpreted as the national diabetes prevalence among U.S. adults, as survey weights were not used in the predictive modelling analysis.

Table 1. Analytic sample sizes

Sample setting	Sample size	Diabetes cases	Diabetes proportion
Non-laboratory model	5,728	1,071	18.7%
Laboratory-enhanced model	5,696	1,039	18.2%

3.2. Model performance

Table 2 reports testing-set performance for the two models in both model settings. In the non-laboratory setting, Random Forest had an AUC of 0.778, accuracy of 0.814, sensitivity of 0.115, specificity of 0.974, and F1-score of 0.188. XGBoost achieved a higher AUC of 0.805 and a much higher sensitivity of 0.791, although its accuracy and specificity were lower. Figure 2 visualizes the same performance metrics across models. The contrast between accuracy and sensitivity is especially clear in the non-laboratory models, where Random Forest achieved high accuracy but very low sensitivity. In comparison, XGBoost showed a more balanced screening performance, with substantially higher sensitivity and a higher AUC.

In the laboratory-enhanced setting, both models improved substantially. Random Forest reached an AUC of 0.964 and XGBoost reached an AUC of 0.968. Sensitivity also improved to 0.785 for Random Forest and 0.875 for XGBoost. These results show that HbA1c added strong classification information.

Table 2. Testing-set model performance

Model setting	Model	AUC	Accuracy	Sensitivity	Specificity	F1-score
Non-laboratory	Random Forest	0.778	0.814	0.115	0.974	0.188
Non-laboratory	XGBoost	0.805	0.700	0.791	0.679	0.496
Laboratory-enhanced	Random Forest	0.964	0.956	0.785	0.994	0.865
Laboratory-enhanced	XGBoost	0.968	0.927	0.875	0.939	0.814

DOI: 10.54254/2753-8818/2026.35377

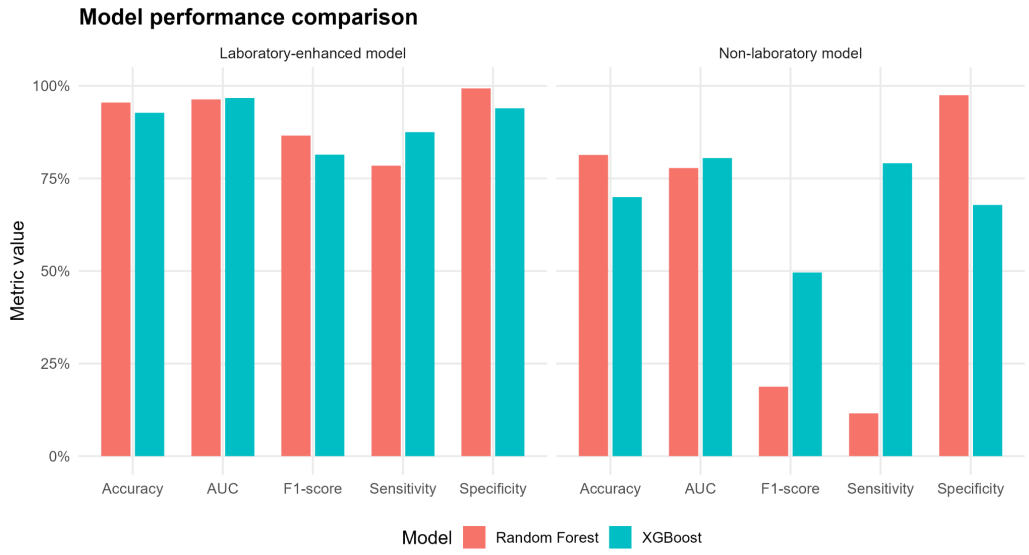

Figure 2. Model performance comparison across the main and laboratory-enhanced settings

3.3. Predictor importance

Figure 3 presents the XGBoost variable importance results for the main non-laboratory model. Age had the largest contribution, followed by waist circumference, diastolic blood pressure, BMI, and systolic blood pressure. Other variables, such as race/ethnicity, income-to-poverty ratio, physical activity, sedentary time, education, and sex, contributed less. These findings indicate that the model primarily relied on age, obesity-related indicators and blood pressure metrics in the absence of laboratory biomarkers. Because BMI and waist circumference are highly related measures of body size, their individual importance rankings should not be interpreted too rigidly.

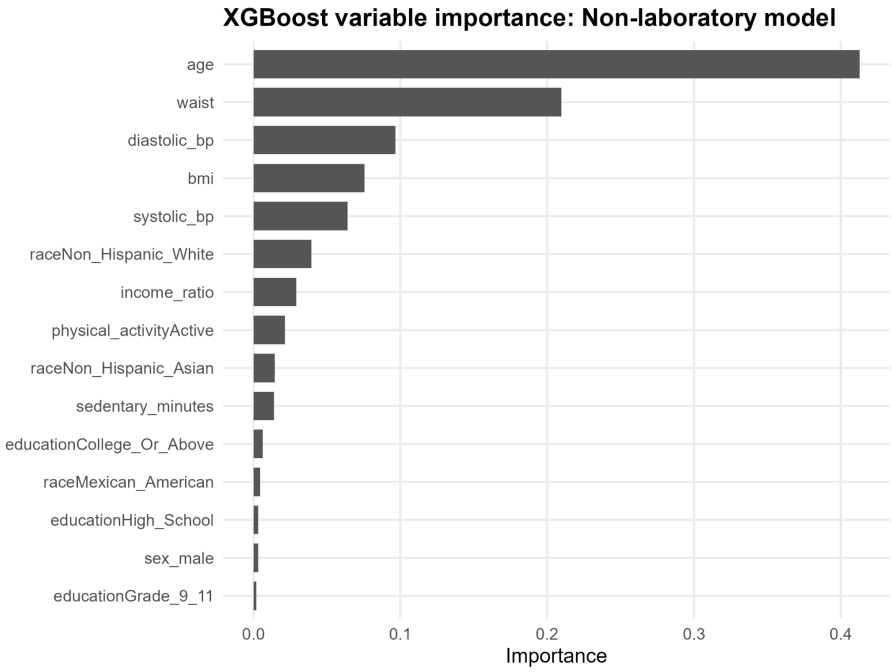

Figure 3. XGBoost variable importance in the non-laboratory model

DOI: 10.54254/2753-8818/2026.35377

4. Discussion

This study examined whether non-laboratory variables can support adult diabetes risk screening and compared two machine learning models in the same prediction task. The results support three main conclusions. First, non-laboratory variables contain meaningful screening information. In the main setting, XGBoost achieved an AUC of 0.805, suggesting that easily collected variables can distinguish diabetes and non-diabetes cases better than chance and at a level that may be useful for preliminary screening.

Second, model evaluation depends strongly on the chosen metric. Random Forest had high accuracy and specificity in the non-laboratory setting, but its sensitivity was only 0.115. This means that it correctly classified many non-diabetes participants but missed most diabetes cases. In a screening context, such a model is less useful despite its high accuracy. XGBoost produced lower overall accuracy, but it achieved much higher sensitivity and a higher F1-score. This outcome reveals that XGBoost is more suitable for screening tasks, as it can identify more diabetic cases at the expense of a higher false positive rate.Third, the inclusion of HbA1c markedly enhanced the overall classification performance. After HbA1c was added, AUC values increased to 0.964 for Random Forest and 0.968 for XGBoost. This result should not be interpreted as evidence that machine learning can replace laboratory testing. Instead, it shows the opposite: laboratory biomarkers provide strong diagnostic information that non-laboratory variables cannot fully reproduce. The supplementary analysis therefore clarifies the role of the two model settings. Non-laboratory models may help identify individuals who should receive further testing, while laboratory-enhanced models reflect the additional information available when clinical biomarkers are measured.

The variable importance results were also consistent with biological and clinical expectations. Age, waist circumference, BMI, and blood pressure variables were among the most important predictors in the XGBoost model. These variables are closely related to metabolic risk and are feasible to collect in low-cost screening settings. However, variable importance does not imply causality. The present study used cross-sectional data, so the results describe classification patterns for current diabetes status, not causal effects or future disease onset.

This analysis has several limitations. First, NHANES is cross-sectional. We cannot predict future diabetes incidence or determine whether predictors caused diabetes. Second, the analysis used complete-case data, which may introduce bias if missingness was systematic. Third, the predictive models did not incorporate sampling weights based on the NHANES survey design. Accordingly, the diabetes proportion in the sample cannot be regarded as an estimate of national prevalence. Fourth, NHANES does not directly classify diabetes type, so the outcome represents adult diabetes status rather than confirmed type 2 diabetes. Future research could use longitudinal cohort data to predict incident diabetes, apply survey-weighted or missing-data methods, and evaluate whether performance differs across sex, age, or race/ethnicity subgroups.

5. Conclusion

Non-laboratory variables can support preliminary adult diabetes risk screening, but their value depends on how model performance is interpreted. In the non-laboratory setting, XGBoost performed better than Random Forest for the screening objective. Random Forest achieved a relatively high accuracy of 0.814 and specificity of 0.974, but its sensitivity was only 0.115, indicating that it correctly excluded most non-diabetes participants but missed a large proportion of diabetes cases. This result shows why accuracy alone is not an adequate evaluation metric in this

DOI: 10.54254/2753-8818/2026.35377

study, especially when diabetes cases represent a smaller part of the analytic sample. In contrast, XGBoost achieved a higher AUC of 0.805 and a much higher sensitivity of 0.791, although its accuracy and specificity were lower. This suggests that XGBoost was more effective at identifying participants with diabetes and was therefore more appropriate for a preliminary screening task.

The supplementary laboratory-enhanced analysis further clarified the role of HbA1c. After HbA1c was added, model performance improved substantially. Random Forest reached an AUC of 0.964, accuracy of 0.956, and specificity of 0.994, while XGBoost reached an AUC of 0.968 and sensitivity of 0.875. These results demonstrate that laboratory biomarkers provide strong additional classification information. Therefore, non-laboratory variables may help identify individuals who should be recommended for further testing, but they cannot replace laboratory-based assessment for accurate diabetes classification.

Overall, this study supports a practical distinction between low-cost screening and clinical diagnosis. Non-laboratory machine learning models, especially XGBoost, may be useful as early screening tools when laboratory data are unavailable. However, HbA1c remains essential for more accurate classification. Because NHANES is cross-sectional, these findings should be interpreted as classification of current diabetes status rather than causal evidence or prediction of future diabetes incidence.

References

[1] International Diabetes Federation. (2025). *IDF Diabetes Atlas: Diabetes facts and figures*. https://idf.org/about-diabetes/diabetes-facts-figures/

[2] Gwira, J. A., Fryar, C. D., & Gu, Q. (2024). Prevalence of total, diagnosed, and undiagnosed diabetes in adults: United States, August 2021–August 2023 (NCHS Data Brief No. 516). National Center for Health Statistics.

[3] Dinh, A., Miertschin, S., Young, A., & Mohanty, S. D. (2019). A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. *BMC Medical Informatics and Decision Making, 19*, Article 211.

[4] Centers for Disease Control and Prevention, National Center for Health Statistics. (n.d.). National Health and Nutrition Examination Survey. National Center for Health Statistics. https://www.cdc.gov/nchs/nhanes/

[5] Akinbami, L. J., Chen, T.-C., Davy, O., Ogden, C. L., Fink, S., Clark, J., Riddles, M. K., & Mohadjer, L. K. (2022). National Health and Nutrition Examination Survey, 2017–March 2020 prepandemic file: Sample design, estimation, and analytic guidelines. *Vital and Health Statistics, 2*(190).

[6] American Diabetes Association Professional Practice Committee. (2026). Diagnosis and classification of diabetes: Standards of Care in Diabetes—2026. *Diabetes Care, 49*(Suppl. 1), S27–S49.

[7] Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32.

[8] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). Association for Computing Machinery.