

Explainable Machine Learning for Diabetes Risk Prediction: Development of a Baseline Predictive Engine for the Wo'tosuga Digital Health Platform Targeting African Populations

EBENEZER A. MARADESA

maradebenazer@gmail.com

University of Abuja <https://orcid.org/0009-0002-8458-6178>

Research Article

Keywords: diabetes mellitus, machine learning, risk prediction, explainable AI, SHAP, LightGBM, Africa, digital health, clinical decision support, BRFSS

Posted Date: July 8th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10241635/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Abstract

Background

Diabetes mellitus poses a growing public health burden across sub-Saharan Africa, with an estimated 54% of affected individuals undiagnosed. Early identification of at-risk individuals is critical to reducing complications, yet laboratory-based screening remains inaccessible to much of the population. Machine learning approaches offer a scalable, non-invasive alternative for population-level risk stratification; however, existing models are predominantly trained on non-African datasets, limiting their direct applicability to African healthcare contexts. This study aimed to develop and evaluate an explainable machine learning model for binary diabetes risk prediction as the foundational predictive engine of the Wo'tosuga digital health platform.

Methods

Data were obtained from the United States Behavioral Risk Factor Surveillance System (BRFSS) across three survey cycles (2019, 2021, 2023), yielding an analytical dataset of 869,319 respondents across 21 predictive features following preprocessing. Pre-diabetic cases were reclassified as diabetic to create a binary outcome reflecting the screening priority of identifying all at-risk individuals. Eight machine learning algorithms were evaluated with sensitivity designated as the primary evaluation criterion. Hyperparameter optimisation was conducted via GridSearchCV with 3-fold cross-validation, with recall as the optimisation metric. SHAP (SHapley Additive exPlanations) values were computed to characterise feature importance.

Results

LightGBM demonstrated the strongest overall performance, achieving a sensitivity of 0.807, specificity of 0.694, AUC-ROC of 0.824, and negative predictive value (NPV) of 0.912. Following hyperparameter optimisation and full-data retraining, the final LightGBM model achieved sensitivity of 0.809, AUC-ROC of 0.828, and NPV of 0.952, representing marginal but consistent improvement. XGBoost (tuned: sensitivity 0.806, AUC 0.824) and logistic regression (sensitivity 0.779, AUC 0.820) served as comparators. SHAP analysis identified age, self-reported general health, BMI, blood pressure status, and time since last medical checkup as the five most influential predictors, consistent across both the primary and baseline models.

Conclusions

A LightGBM model trained on population-level survey data demonstrates promising discrimination performance for diabetes risk screening, with a strong NPV supporting its use as a triage instrument. The exclusive use of non-African training data is the primary limitation and the central motivation for the Wo'tosuga platform design, which incorporates an anonymous data donation mechanism to collect African-specific data for model retraining. These results establish the Phase 1 performance baseline for a planned programme of African-context model development and validation.

1. INTRODUCTION

Diabetes mellitus is one of the fastest-growing non-communicable diseases globally, affecting an estimated 537 million adults as of 2021 and projected to reach 783 million by 2045.¹ The African continent is disproportionately affected: the Africa region is expected to see the greatest relative increase in diabetes prevalence of any world region between 2021 and 2045.¹ Of particular concern is the high proportion of undiagnosed cases, estimated at 54% across Africa, reflecting limited penetration of routine screening and significant healthcare access barriers.² Undiagnosed diabetes allows progression of microvascular and macrovascular complications that are costly and largely preventable with early intervention.

Conventional laboratory-based diabetes screening, relying on fasting plasma glucose, oral glucose tolerance testing, or glycated haemoglobin measurement, is inaccessible to a large proportion of the African population due to infrastructural, financial, and geographic constraints. There is consequently an urgent need for scalable, low-cost screening instruments capable of deployment in resource-limited settings, where a missed diagnosis may never be corrected by an alternative clinical pathway.

Machine learning (ML) methods have demonstrated considerable promise for diabetes risk prediction from non-invasive, self-reported variables, offering a complement to laboratory-based screening.^{3,4} Gradient boosting algorithms, in particular, have consistently achieved strong discriminative performance on population health datasets, with reported AUC-ROC values typically ranging from 0.78 to 0.88.^{5,6} The integration of explainability frameworks such as SHAP (SHapley Additive exPlanations)⁷ has further enhanced the clinical utility of these models by providing interpretable per-feature contributions that support clinical decision-making and patient communication.

However, the vast majority of published diabetes prediction models are trained on datasets from North American or European populations, most notably the Pima Indian Diabetes Dataset and the CDC Behavioral Risk Factor Surveillance System (BRFSS).^{3,4,5} The generalisability of these models to African populations is substantially limited by differences in the distribution of risk factors, healthcare access patterns, and the phenotypic characteristics of diabetes in African contexts.⁸ The most directly comparable recent study, the AWI-Gen cohort study from northern Ghana,⁸ used purpose-collected African cohort data and achieved AUC-ROC of 0.845 with sensitivity of 78.2%, but remains a single-site study from a specific subregion of West Africa. Population-level African diabetes data of the scale and breadth available from BRFSS-equivalent surveillance systems does not currently exist for sub-Saharan Africa.

A second critical limitation of the existing literature is the predominant use of overall accuracy as the primary evaluation metric, which is inappropriate for imbalanced screening tasks. For a diabetes screening tool intended for deployment in low-access settings, sensitivity, the proportion of true diabetics correctly identified, is the clinically relevant metric.⁵ A model that achieves high accuracy by predicting the majority class provides no clinical utility as a screening instrument. This issue has been noted in the literature but inconsistently addressed in practice.³

This paper presents the Phase 1 results of Wo'tosuga, a digital health platform designed to address the diabetes burden in Africa through a dual research-and-clinical-product architecture. The specific objectives of this study were: (1) to develop and evaluate a machine learning model for binary diabetes risk classification using BRFSS population health survey data, with sensitivity as the primary metric; (2) to compare eight candidate algorithms and identify the most suitable baseline model; (3) to characterise feature importance using SHAP analysis and

assess clinical interpretability; and (4) to establish a reproducible baseline for subsequent African-context model development, supported by anonymous clinical data collection through the Wo'tosuga web application.

2. RELATED WORK

2.1 Machine learning for diabetes risk prediction

The application of machine learning to diabetes risk prediction has grown substantially over the past decade, with gradient boosting methods consistently achieving the strongest discriminative performance on tabular population health data. Jiang et al.⁵ applied LightGBM to 2021 BRFSS survey data with feature enumeration and achieved robust predictive performance, identifying BMI, age, and general health as primary predictors, a finding directly consistent with the results of this study. Allani evaluated multiple classifiers including LightGBM and XGBoost on BRFSS data with SHAP and LIME interpretability, deployed as an interactive web application, representing the closest methodological comparator to the current work.⁶ These studies demonstrate that BRFSS-based diabetes prediction with gradient boosting and SHAP explainability has become an established, if crowded, methodological territory.

A 33-year bibliometric and systematic review of ML for type 2 diabetes prediction published in 2025 confirmed that Random Forest, SVM, XGBoost, and KNN remain the dominant algorithms in the field, with SHAP emerging as the standard interpretability framework.³ The review highlighted persistent concerns about data leakage, class imbalance handling, and overfitting as the principal methodological weaknesses limiting clinical deployment, concerns directly addressed in the methodology of the present study.

A Tennessee BRFSS-specific study applied explainable ML models to 2023 BRFSS data with SHAP analysis, targeting diabetes classification among US adults, the most geographically and methodologically proximate comparator to this work, differing primarily in its single-state scope and the absence of an explicit deployment pathway or African-context framing.⁴

2.2 African-context diabetes prediction

Studies specifically addressing diabetes prediction in African populations are substantially fewer and more recent than the global literature. The most directly comparable work is Adda et al.⁸, who applied eight ML classifiers to data from 2,010 participants in the H3Africa AWI-Gen cohort in northern Ghana. Using Bayesian hyperparameter optimisation, SMOTE for class imbalance, and stratified 5-fold cross-validation, their optimised XGBoost model achieved AUC-ROC of 0.845 and sensitivity of 78.2%. Critically, the study excluded blood glucose as a predictor to avoid target leakage, and noted that its inclusion increased AUC by 11.5%, a methodological decision also adopted implicitly in the present study, which relies entirely on self-reported variables. The AWI-Gen cohort is private and purpose-collected, with a feature set emphasising clinical anthropometric and metabolic measurements rather than the behavioural and socioeconomic variables available in BRFSS, representing a complementary rather than competing approach.

The relative scarcity of African-specific models reflects the well-documented African medical data gap: large-scale population health surveillance data of the type routinely collected in the United States through BRFSS does not currently exist for most sub-Saharan African countries. This gap is both the primary limitation of the present study and the central motivation for the Wo'tosuga platform design.

2.3 Explainability in clinical machine learning

SHAP, introduced by Lundberg and Lee,⁷ has become the dominant framework for post-hoc explainability of tree-based models in clinical applications, providing theoretically grounded per-feature attributions based on Shapley values from cooperative game theory. TreeSHAP, the efficient implementation for tree ensembles, enables exact Shapley value computation in polynomial time, making it practical for large-scale datasets.⁷ In clinical decision support, SHAP is valued both for global feature importance (understanding which variables drive population-level predictions) and for local explanations (explaining individual patient predictions to clinicians or patients). The present study implements both levels of explanation, with local SHAP values forming the basis for the personalised risk report in the Wo'tosuga web application.

2.4 Positioning of the present study

The present study differs from existing BRFSS-based diabetes prediction work in three principal respects. First, it adopts sensitivity rather than accuracy or AUC-ROC as the primary evaluation criterion, reflecting an explicit clinical design rationale for screening in low-access settings. Second, it positions the model not as a standalone publication but as the validated baseline engine for a deployed digital health platform, with the SHAP explainability layer integrated directly into the application rather than presented only as a research output. Third, and most distinctively, it frames the use of non-African training data not as a limitation to minimise but as the foundational motivation for the platform's data collection architecture, which is designed to progressively close the African data gap through anonymous user data donation.

3. METHODS

3.1 Data source and study population

Data were obtained from the Behavioral Risk Factor Surveillance System (BRFSS), an annual cross-sectional telephone survey administered by the United States Centers for Disease Control and Prevention (CDC) to assess health-related behaviours, chronic conditions, and preventive service use among non-institutionalised adults aged 18 years and older. Three survey cycles were utilised: 2019, 2021, and 2023. These cycles were selected on the basis of consistent availability of self-reported hypertension status, which is collected as part of the BRFSS rotating core module on a biennial basis and was identified in preliminary analysis as a clinically significant predictor. The 2017 and 2020 cycles were excluded due to the absence of hypertension data; the 2022 and 2024 cycles were similarly excluded for the same reason, resulting in a balanced selection of three alternate-year cycles spanning five years.

The combined raw dataset prior to cleaning comprised approximately 1,250,000 respondents across the three cycles. Minor naming inconsistencies in income-related variables across survey cycles were harmonised prior to concatenation.

3.2 Data preprocessing and feature engineering

Variables with greater than 50% missing values were removed prior to analysis, including average alcoholic drinks per day, which was excluded on this basis. Complete-case analysis was subsequently applied to the remaining variables, yielding a final analytical dataset of 869,319 respondents across 21 predictive features. This

approach may introduce selection bias toward respondents with more complete health records; this limitation is explicitly acknowledged.

Two categorical variables were recoded to improve clinical interpretability and reduce dimensionality. Marital status was collapsed from six categories to three: Married/Cohabiting (married; member of an unmarried couple), Previously Married (divorced; separated; widowed), and Single (never married). Employment status was collapsed from eight categories to six: Employed (employed for wages; self-employed), Retired, Unemployed (out of work), Not in Workforce (homemaker), Student, and Unable to Work.

A socioeconomic tier variable was derived from BRFSS income bracket data by mapping respondents into three categories, Low (difficulty affording basic necessities), Middle (basic necessities met with some surplus), and High (substantial surplus), to provide a transferable socioeconomic indicator applicable to future data collection from non-US populations where dollar-denominated income brackets would be uninformative.

The diabetes outcome variable was recoded as a binary variable. Pre-diabetic cases were reclassified as diabetic (positive class, label = 1), reflecting the clinical rationale that pre-diabetes requires active intervention and that failing to flag a pre-diabetic individual in a screening context represents a consequential clinical error. The final class distribution was 734,009 non-diabetic (84.4%) and 135,310 diabetic (15.6%).

Yes/No and True/False columns were recoded as 0/1 for easier computation. Afterwards, preprocessing was implemented as a scikit-learn ColumnTransformer pipeline applying StandardScaler to continuous numeric features (age, BMI, education, income, general health rating, physical health days, mental health days, smoking history, alcohol consumption, stroke history, heart attack history, coronary heart disease history, healthcare cost barrier, checkup recency, personal doctor access, socioeconomic tier) and OneHotEncoder with `handle_unknown='ignore'` to categorical features (sex, marital status, employment status, exercise status, blood pressure status).⁹

3.3 Model development and evaluation

Eight machine learning algorithms were evaluated: logistic regression (LR), decision tree (DT), extra trees (ET), gradient boosting (GB), AdaBoost (AB), XGBoost (XGB), LightGBM (LGBM), and K-nearest neighbours (KNN). All models were implemented within a unified scikit-learn pipeline framework.⁹ Class imbalance was addressed through cost-sensitive learning: logistic regression, decision tree, extra trees, AdaBoost, and LightGBM utilised the `class_weight='balanced'` parameter; XGBoost utilised `scale_pos_weight` set to the ratio of negative to positive class instances ($734,009 / 135,310 \approx 5.42$). This approach was preferred over synthetic oversampling (SMOTE) as it avoids the generation of synthetic minority-class instances and preserves the natural class distribution, maintaining well-calibrated probability estimates (furthermore, subsequent attempts at random oversampling with SMOTE produced no notable improvements in model performances, further justifying the preference for class weighting).

Data were partitioned into training (95%) and test (5%) sets using stratified random sampling to preserve the class distribution. Hyperparameter optimisation was subsequently performed for the three highest-performing models (LightGBM, XGBoost, logistic regression) using GridSearchCV with 3-fold cross-validation on a stratified random subsample of 200,000 respondents, with recall (sensitivity) as the optimisation criterion. The best-performing hyperparameter configurations identified were: LightGBM (`learning_rate = 0.05`, `n_estimators = 300`, `num_leaves = 31`); XGBoost (`learning_rate = 0.01`, `max_depth = 5`, `n_estimators = 300`, `subsample = 0.8`); logistic

regression (C = 0.01, solver=lbfgs). Final models were retrained on the full training dataset using these configurations and evaluated on the held-out test set.

3.4 Evaluation metrics

Model performance was assessed using the following metrics: sensitivity (recall; proportion of true diabetics correctly identified), specificity (proportion of true non-diabetics correctly identified), positive predictive value (PPV), negative predictive value (NPV), F1 score, AUC-ROC, average precision (area under the precision-recall curve), and Brier score (mean squared error of predicted probabilities, reflecting calibration quality). Sensitivity was designated as the primary evaluation metric, consistent with the clinical objective of maximising detection of at-risk individuals in a population screening context where false negatives carry the greatest clinical cost.

3.5 Explainability analysis

SHAP values were computed for the primary LightGBM model using the TreeExplainer algorithm,⁷ which provides exact Shapley values for tree-based models. Global feature importance was characterised by mean absolute SHAP value across test set predictions. Dependence plots were generated for the two most important continuous features (age and BMI) to characterise the direction and shape of their associations with predicted risk. All SHAP analyses were performed on a random subsample of 5,000 test set instances for computational efficiency.

3.6 Reproducibility

All analyses were conducted in Python 3.11 using scikit-learn,⁹ XGBoost,¹⁰ LightGBM,¹¹ SHAP,⁷ pandas, and matplotlib. A random seed of 42 was used throughout all stochastic operations. The complete analysis pipeline and pre-trained model are available at <https://github.com/maradeben/wotosuga-core/>. To ensure reproducibility and maintain a transparent development workflow, Git was utilized for version control throughout the codebase and model development lifecycle.

Portions of this manuscript and certain code snippets were drafted with the assistance of Claude (Anthropic, version Sonnet 4.5), a large language model. All analytical decisions, model implementation, clinical reasoning, interpretation of results, and final manuscript content were reviewed, verified, and approved by the author. The LLM was not involved in execution of data analysis or code.

4. RESULTS

4.1 Dataset characteristics

Following harmonisation, preprocessing, and complete-case analysis, the final analytical dataset comprised 869,319 respondents from the 2019, 2021, and 2023 BRFSS cycles. The dataset included 21 predictive features spanning demographic characteristics, anthropometric measurements, health behaviours, comorbid conditions, and healthcare access variables. The binary diabetes outcome comprised 734,009 non-diabetic respondents (84.4%) and 135,310 diabetic respondents (15.6%), representing an approximately 5.4:1 class imbalance.

4.2 Model comparison

Table 1 presents performance metrics for all eight evaluated algorithms on the held-out test set. Two distinct performance profiles were observed across the candidate models. LightGBM, XGBoost, and logistic regression

achieved high sensitivity (0.779–0.808) with moderate specificity (0.701–0.719) and AUC-ROC values ranging from 0.820 to 0.828. The remaining five models, gradient boosting, AdaBoost, extra trees, bagging, KNN, and decision tree, achieved substantially lower sensitivity (0.190–0.305) despite higher accuracy, reflecting a systematic bias toward predicting the majority (non-diabetic) class. These models, while appearing favourable on accuracy metrics, are clinically unsuitable as screening instruments because they fail to identify the majority of at-risk individuals.

Table 1: Performance metrics for eight candidate algorithms on the held-out test set ($n = 43,466$). * denotes best value for that metric across all models. † Primary model selected for hyperparameter optimisation and deployment. Models ordered by sensitivity. PPV = positive predictive value; NPV = negative predictive value; AUC-ROC = area under the receiver operating characteristic curve; Avg Prec. = average precision (area under precision-recall curve).

Model	Accuracy	Sensitivity	Specificity	PPV	NPV	F1	AUC-ROC	Avg Prec.	Brier
LightGBM†	0.718	0.808*	0.701	0.333	0.952*	0.471	0.828*	0.458*	0.177
XGBoost	0.719	0.799	0.705	0.333	0.950	0.470	0.826	0.456	0.176
Logistic Regression	0.728	0.779	0.719	0.338	0.946	0.471	0.820	0.432	0.179
Gradient Boosting	0.852	0.190	0.974	0.575	0.867	0.285	0.826	0.455	0.105
AdaBoost	0.850	0.210	0.968	0.550	0.869	0.304	0.821	0.443	0.154
Bagging	0.835	0.204	0.951	0.434	0.866	0.278	0.753	0.330	0.124
KNN	0.830	0.242	0.938	0.418	0.870	0.306	0.730	0.308	0.126
Decision Tree	0.785	0.305	0.874	0.308	0.872	0.307	0.589	0.203	0.214

Of particular note is the performance of gradient boosting, which achieved an AUC-ROC of 0.826, virtually identical to LightGBM, while achieving a sensitivity of only 0.190. This finding illustrates the critical importance of metric selection: a model that would appear competitive or superior under accuracy or AUC-ranked comparison is revealed as clinically unsuitable when evaluated by the primary criterion of sensitivity. The near-equivalence of logistic regression, XGBoost, and LightGBM in the AUC-ROC range 0.820–0.828 further suggests that most of the predictive signal in this dataset is captured by a relatively small set of linearly-separable risk factors, with ensemble methods providing modest incremental gains.

4.3 Hyperparameter optimisation and full-data retraining

Table 2 presents the performance comparison between baseline and tuned models following GridSearchCV optimisation and retraining on the full training dataset. Tuning produced marginal improvements in sensitivity for LightGBM and XGBoost, confirming the robustness of both models across hyperparameter configurations.

Table 2: Baseline versus tuned model performance for the three highest-performing algorithms. Best hyperparameter configurations: LightGBM ($learning_rate = 0.05$, $n_estimators = 300$, $num_leaves = 31$); XGBoost ($learning_rate = 0.01$, $max_depth = 5$, $n_estimators = 300$, $subsample = 0.8$); logistic regression ($C = 0.01$,

solver=lbgfs). GridSearchCV conducted on a stratified subsample of 200,000 respondents; final models retrained on full training data (825,853 respondents).

Model	Configuration	Sensitivity	Specificity	PPV	NPV	AUC-ROC	Brier
LightGBM	Baseline	0.807	0.694	0.328	0.950	0.824	0.180
LightGBM	Tuned (full)	0.809	0.701	0.333	0.952	0.828	0.177
XGBoost	Baseline	0.799	0.705	0.333	0.950	0.826	0.176
XGBoost	Tuned (full)	0.806	0.696	0.329	0.951	0.824	0.179
Logistic Regression	Baseline	0.786	0.709	0.333	0.947	0.819	0.182
Logistic Regression	Tuned (full)	0.779	0.719	0.338	0.946	0.820	0.179

4.4 SHAP feature importance

Figure 1 presents the mean absolute SHAP values for the top 20 features from the final LightGBM model. Age (SHAP importance: 0.538), self-reported general health status (0.487), and BMI (0.368) were the three most influential predictors. Blood pressure status (normal/no: 0.314; hypertension/yes: 0.209) collectively ranked fourth, reflecting the substantial contribution of hypertension as a diabetes risk predictor. Time since last medical checkup (0.252) and alcohol consumption (0.191) completed the top seven features. Features with near-zero importance, including socioeconomic tier (0.00005) and the binary exercise variable (0.00003), contributed negligibly to model predictions.

A high degree of structural consensus was observed between the top absolute SHAP values from the LGBM, and the feature importances of the best Logistic Regression model, which are presented in Table 3. Specifically, the models demonstrated identical rank-order positioning for the top five (5) predictors: Age, General Health Self-Rating, Body Mass Index (BMI), Normal Blood Pressure and Time Since Last Medical Checkup. Furthermore, within the top five features, both frameworks converged on the exact same clinical and behavioral dimensions, with hypertension status and medical checkup recency rounding out the primary risk tier. This cross-architecture convergence indicates that the model's predictive weightings are stable across differing mathematical frameworks.

Table 3
Feature importances of best Logistic Regression model

Feature Name	Coefficient
Age	0.5682
General Health Self-Rating	-0.5102
Body Mass Index (BMI)	0.4552
Blood Pressure: Normal/No	-0.4482
Time Since Last Medical Checkup	-0.3577

Figure 2 presents the SHAP beeswarm plot for the top 20 features, showing the direction and distribution of feature effects across test set predictions. The beeswarm plot reveals several clinically interpretable patterns. For age, SHAP values increase monotonically with age, with younger respondents showing strongly negative SHAP contributions (protective) and older respondents showing strongly positive contributions (risk-increasing), confirming a clear age-dependent risk gradient. For BMI, the dependence plot (Fig. 3) reveals a non-linear relationship: SHAP values increase steeply as BMI rises from the underweight range, plateau at moderate values in the normal-to-overweight range, and rise again sharply above approximately BMI 25–30, consistent with the obesity threshold for metabolic risk. The blood pressure feature (normal/no) shows a strong negative SHAP contribution for respondents without hypertension and a positive contribution for those with it, confirming hypertension as a risk-amplifying factor. General health self-rating shows an inverse relationship, with respondents rating their health as poor or fair showing positive SHAP contributions and those rating it as good or excellent showing negative contributions.

The direction and approximate rank ordering of feature importance was consistent between the LightGBM SHAP values and the logistic regression coefficient magnitudes, with age, general health, BMI, blood pressure, and checkup recency occupying the top five positions in both models. This cross-model consistency provides evidence that the identified associations reflect stable, generalisable relationships between clinical and behavioural variables and diabetes risk, rather than model-specific artefacts.

5. DISCUSSION

5.1 Principal findings

This study developed and evaluated a machine learning-based diabetes risk prediction model using population-level survey data from three BRFSS cycles (2019, 2021, 2023), establishing the Phase 1 performance baseline for the Wo'tosuga digital health platform. Among eight candidate algorithms, LightGBM demonstrated the strongest overall performance, achieving sensitivity of 0.809, AUC-ROC of 0.828, and NPV of 0.952. These results are consistent with published diabetes prediction models on population health datasets, where AUC-ROC values of 0.78–0.88 are typically reported,^{3,5,6} and represent a competitive performance baseline given the exclusive reliance on self-reported, non-laboratory variables.

The strong NPV of 0.952 is of particular clinical relevance. For a screening tool intended for deployment in settings where follow-up diagnostic capacity is limited, the reliability of a negative result is as important as the sensitivity of a positive one. A NPV of 0.952 indicates that 95 of every 100 individuals cleared by the model are genuinely non-diabetic, providing a high degree of confidence in negative predictions and substantially limiting the risk of false reassurance, the primary clinical hazard of a screening instrument in a low-access setting.

The observed near-equivalence of logistic regression, XGBoost, and LightGBM across AUC-ROC (0.820–0.828) and sensitivity (0.779–0.809) is itself a meaningful finding. It suggests that the predictive signal in this dataset is largely captured by a small set of clinically intuitive, approximately linearly-separable risk factors, with ensemble methods providing only marginal incremental gains. This observation is consistent with published literature on BRFSS-based diabetes prediction,^{5,6} and has practical implications: a logistic regression model, simpler, faster, and more easily deployed in low-resource settings, achieves competitive performance as a potential alternative in contexts where computational resources are severely constrained.

5.2 Feature importance and clinical interpretability

The consistent identification of age, BMI, general health self-rating, hypertension status, and checkup recency as the five most important predictors across both the LightGBM and logistic regression models provides strong evidence that the model has captured genuine biological and socioeconomic determinants of diabetes risk. Each of these associations has robust epidemiological support. Advancing age and elevated BMI are among the most strongly established modifiable and non-modifiable risk factors for type 2 diabetes.¹ Self-reported general health is a validated proxy for cumulative disease burden with established predictive validity for objective health outcomes. The shared pathophysiological mechanisms between hypertension and diabetes, including insulin resistance, endothelial dysfunction, and common upstream metabolic disturbances, explain the strong predictive contribution of blood pressure status.¹

The negative association observed between alcohol consumption and diabetes risk warrants careful interpretation. The observed direction is likely a surveillance artefact: in BRFSS data, individuals with established chronic illness, including diagnosed diabetes, systematically report lower alcohol consumption, creating a spurious protective association.⁵ This finding should not be interpreted as evidence of a protective effect of alcohol and is explicitly flagged as a limitation of self-reported cross-sectional data.

Checkup recency showed a negative association with diabetes risk, meaning more recent healthcare contact was associated with higher predicted risk, reflecting ascertainment bias rather than a causal relationship. Individuals with established or suspected chronic conditions are more likely to have had recent medical contact, making checkup recency a proxy for existing disease burden rather than a causal risk factor. This variable should be interpreted accordingly in clinical deployment.

The near-zero SHAP importance of the socioeconomic tier variable is noteworthy given its theoretical relevance to diabetes risk through pathways of food security, healthcare access, and chronic stress. This may reflect the coarseness of the three-category classification, the indirect nature of the BRFSS income variable from which it was derived, or genuine attenuation of the socioeconomic signal by the stronger biological predictors in the model. Future iterations using more granular socioeconomic data, or data from populations with greater socioeconomic heterogeneity, may reveal stronger associations.

5.3 The African data gap and the Wo'tosuga research loop

The exclusive use of BRFSS data, a United States population survey, as the training dataset is the primary limitation of the current model and motivates the central design feature of the Wo'tosuga platform. African populations differ from North American populations in the distribution of diabetes risk factors, including the higher prevalence of lean phenotype diabetes, differential patterns of healthcare utilisation, and distinct nutritional and lifestyle profiles.⁸ A model trained exclusively on US survey data cannot be assumed to generalise to African populations without empirical validation, and direct comparison with the AWI-Gen cohort results,⁸ which achieved AUC-ROC of 0.845 on purpose-collected African data, suggests that African-specific data may yield meaningfully stronger performance.

This limitation is, however, intentionally framed as the foundational research question rather than a constraint to minimise. The Wo'tosuga web application will incorporate an optional, approved anonymous data donation mechanism at the point of risk prediction, through which consenting users contribute their risk factor inputs and

predicted classification to a growing African dataset. This dataset will be used to retrain and validate the model in subsequent development phases. The platform thus functions simultaneously as a clinical screening instrument and as a data collection instrument, creating a self-improving research loop designed to progressively close the African data gap. This approach transforms a methodological limitation into a translational research pathway, the US-trained baseline model enables the platform to be deployed, and platform deployment generates the African data needed to improve the model.

Phase 2 of the research programme will evaluate the transferability of the BRFSS-trained model to African populations as African-specific data accumulates, addressing the question of how much performance is lost when a US-trained model is applied to African respondents and how much African data is required before retraining produces meaningful improvements. These are the genuinely novel empirical questions that the current Phase 1 establishes the infrastructure to answer.

5.4 Limitations

Several limitations merit explicit acknowledgement. First, complete-case analysis following removal of high-missingness variables introduces potential selection bias toward respondents with more complete health records, who may differ systematically from those with incomplete responses in ways relevant to diabetes risk. Second, all predictor variables are self-reported and subject to recall bias and social desirability effects. Third, the cross-sectional design precludes causal inference and longitudinal risk characterisation. Fourth, the non-African training population limits direct transferability to African clinical contexts. Fifth, coronary heart disease was retained in the feature set despite lower reliability of self-report in low-resource settings (where the underlying diagnosis typically requires imaging or catheterisation), representing an inconsistency with the clinical reasoning that guided exclusion of other low-reliability variables; this feature should be revisited in the next model iteration. Sixth, the socioeconomic tier variable, while designed for cross-population transferability, contributed negligible predictive signal in the current model and may require reconceptualisation. Finally, no external validation has been performed; all reported metrics reflect internal validation on a held-out subset of the BRFSS data.

5.5 Clinical deployment considerations

The Wo'tosuga model is designed as a screening instrument rather than a diagnostic tool. A positive prediction should prompt clinical follow-up including laboratory-based glucose assessment, not substitute for it. In African primary care contexts, the model is most appropriately deployed as a triage instrument to prioritise individuals for further evaluation, particularly in settings where laboratory capacity is constrained. The strong NPV supports its use in this role: a negative result provides reliable reassurance, while a positive result appropriately escalates care. The SHAP-based individual explanation layer, which identifies the top contributing risk factors for each prediction, is an essential component of responsible clinical deployment, supporting patient education and enabling clinician oversight of the prediction rationale.

6. CONCLUSIONS

This study demonstrates that a LightGBM model trained on population-level US survey data achieves promising discriminative performance for diabetes risk screening (sensitivity 0.809, AUC-ROC 0.828, NPV 0.952), with a strong negative predictive value supporting its utility as a triage instrument in low-access settings. The identified feature importance hierarchy is consistent with established diabetes epidemiology and cross-validated against logistic regression coefficients, supporting the clinical credibility of the model's learned associations. The study

establishes the Phase 1 performance baseline for the Wo'tosuga digital health platform and provides the technical infrastructure for subsequent phases of African-context model development. The central contribution of this work is not the predictive model itself, gradient boosting on population health survey data is an established and well-documented approach, but the translational architecture: a deployed screening tool that simultaneously functions as a data collection instrument designed to close the African medical data gap that represents the model's primary limitation.

Declarations

Ethics approval and consent to participate

BRFSS data are publicly available and de-identified; no additional ethical approval was required for this secondary analysis. The Wo'tosuga data donation component will be subject to an appropriate review board approval prior to deployment.

Data Availability

The BRFSS datasets analysed during this study are publicly available from the U.S. Centers for Disease Control and Prevention at <https://www.cdc.gov/brfss/>. The versions used for this study were downloaded from Kaggle at <https://www.kaggle.com/datasets/spandanjit2005/brfss-diabetes-indicator-dataset>. The analysis code, preprocessing pipeline, and pre-trained model will be made available at <https://github.com/maradeben/wotosuga-core/> under an MIT licence upon submission of this manuscript.

Competing interests

The author declares no competing interests.

Funding

This research received no external funding.

Author contributions

EM conceived the study, designed the analytical framework, implemented the complete model pipeline, conducted all statistical analyses, and wrote the manuscript in its entirety.

Consent to participate

Not applicable. This study utilised publicly available, de-identified secondary data from the United States Behavioral Risk Factor Surveillance System (BRFSS). No individual participant identification was possible and no consent to participate was required.

Consent for publication

Not applicable.

References

1. Sun H, Saeedi P, Karuranga S, Pinkepank M, Ogurtsova K, Duncan BB et al (2022) IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. *Diabetes Res Clin Pract* 183:109119. 10.1016/j.diabres.2021.109119
2. Ogurtsova K, Guariguata L, Barengo NC, Ruiz PL, Sacre JW, Karuranga S et al (2022) IDF diabetes Atlas: Global estimates of undiagnosed diabetes in adults for 2021. *Diabetes Res Clin Pract* 183:109118. 10.1016/j.diabres.2021.109118
3. Kiran T, Xie Z, Anjum A, Ball J, Pierscionek B, Russell S (2025) Machine learning and artificial intelligence in type 2 diabetes prediction: a comprehensive 33-year bibliometric and literature analysis. *Front Digit Health* 7:1557467. 10.3389/fdgth.2025.1557467
4. Muhammad MA, Sani J, Ahmed MM (2025) Exploring Explainable Machine Learning for Predicting and Interpreting Self-Reported Diabetes among Tennessee Adults: Insights from the 2023 Behavioral Risk Factor Surveillance System (BRFSS). *J Prim Care Community Health* 16. 10.1177/21501319251400546
5. Jiang L, Yang Z, Liu G, Xia Z, Yang G, Gong H, Wang J, Wang L (2024) A feature optimization study based on a diabetes risk questionnaire. *Front Public Health* 12:1328353. 10.3389/fpubh.2024.1328353
6. Allani U (2025) Interactive Diabetes Risk Prediction Using Explainable Machine Learning: A Dash-Based Approach with SHAP, LIME, and Comorbidity Insights. *arXiv preprint arXiv:2505.05683*
7. Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst* 30:4768–4777
8. Adda RB Improving Type 2 Diabetes Prediction: Comparative Evaluation of Machine Learning Classifiers Using Balanced Data from the AWI-Gen Cohort. *Res Square preprint*. 10.21203/rs.3.rs-8019155/v1. Published 2025 Nov 4.
9. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O et al (2011) Scikit-learn: Machine Learning in Python. *J Mach Learn Res* 12:2825–2830
10. 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*. :785–794. 10.1145/2939672.2939785
11. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu TY (2017) LightGBM: A highly efficient gradient boosting decision tree. *Adv Neural Inf Process Syst* 30:3146–3154
12. Centers for Disease Control and Prevention. Behavioral Risk Factor Surveillance System Survey Data. Atlanta: U.S. Department of Health and Human Services (2019) 2021, 2023. Available from: <https://www.cdc.gov/brfss/>

Figures



Figure 1

SHAP mean absolute feature importance bar chart.

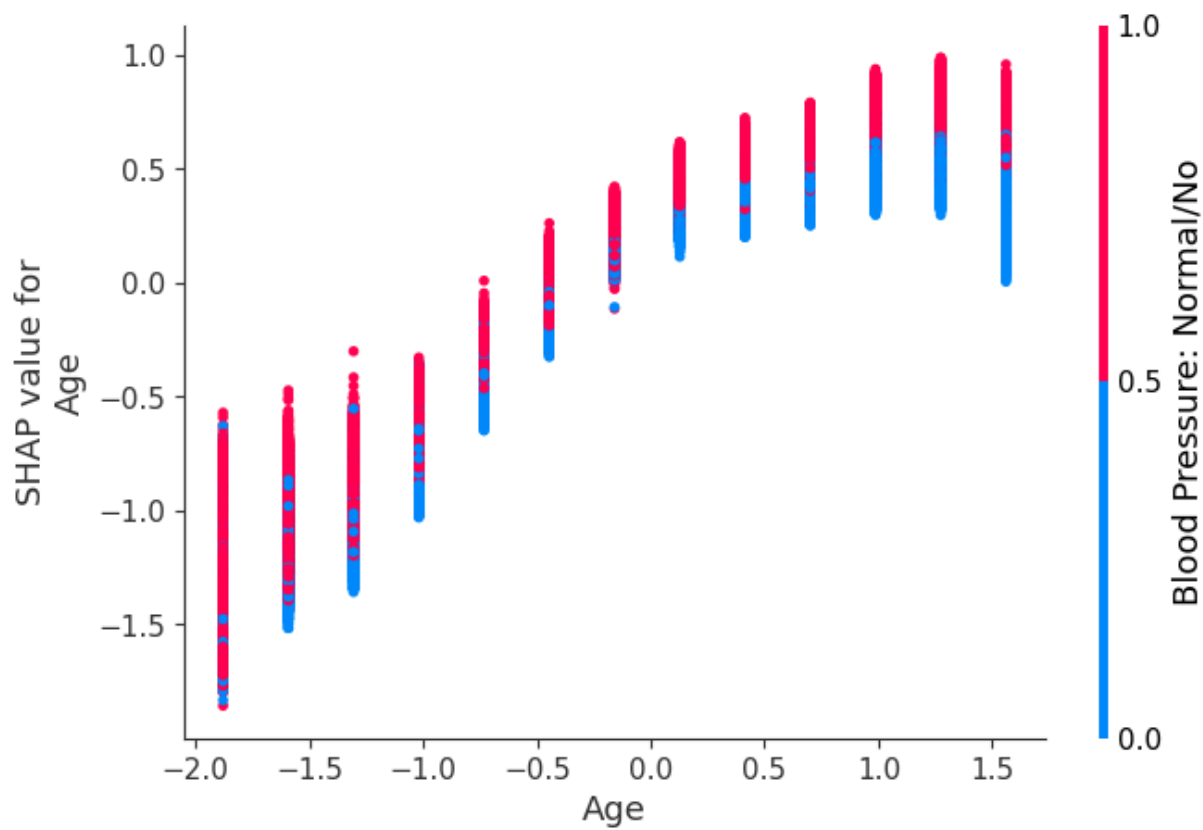


Figure 2

SHAP beeswarm plot (Scaled ages on the horizontal axis).

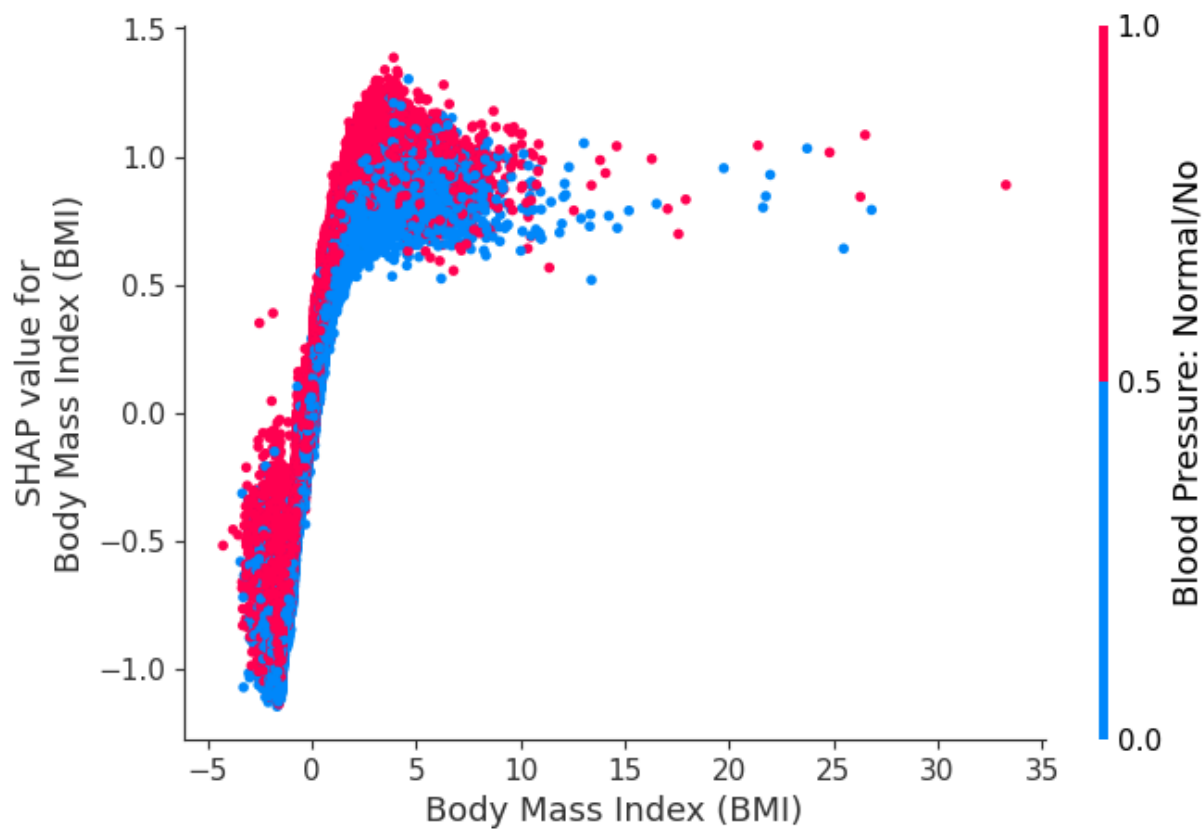


Figure 3

SHAP beeswarm plot (BMI).

