An Explainable Machine Learning Framework for Type 2 Diabetes Risk Prediction: Integrating XGBoost, SHAP/LIME Interpretability, and a Conversational AI Assistant

A Design and Implementation Study on the CDC BRFSS 2023 Survey Dataset

Prepared as part of a Data Science Phase 5 Capstone Project
Diabetes Risk Predictor — Explainable AI (XAI) Platform

Authors
Stephen Mwaura
Angela Masaki
Diana Byego
Kevin Kisengu
Orandi Felix

Institution
Moringa School
Date 2026

Abstract

Type 2 diabetes mellitus (T2DM) affects hundreds of millions of adults worldwide and is frequently undiagnosed until complications arise, making early, low-cost, non-invasive risk screening a critical public health goal. This paper presents the design, implementation, and evaluation of an end-to-end explainable machine learning (XAI) platform for T2DM risk prediction, built on the Centers for Disease Control and Prevention's (CDC) 2023 Behavioral Risk Factor Surveillance System (BRFSS) survey, comprising 429,086 respondents. After extensive exploratory data analysis, feature engineering, and data preprocessing, we trained and evaluated five classification models Logistic Regression, Decision Tree, Random Forest, XGBoost, and LightGBM on a curated set of fourteen clinically and behaviourally relevant features (including body mass index, age, general health, hypertension, cholesterol status, smoking history, physical activity, and comorbidities). The best-performing model, XGBoost, achieved a recall of 79.7% and a receiver operating characteristic area under the curve (ROC-AUC) of 82.9% on held-out test data, prioritising sensitivity to reduce missed at-risk cases. LightGBM delivered comparable performance. To address the opacity of gradient-boosted ensembles, the system integrates SHapley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME) to provide both global (population-level) and local (per-patient) interpretability. Beyond the predictive core, the platform contributes a full-stack deployment: a FastAPI backend exposing single and batch prediction endpoints, a Next.js dashboard with interactive visual analytics, PDF/CSV report generation with email delivery, and "Dida," a large language model (LLM)-based conversational assistant that guides users through risk assessment in natural language while a deterministic client-side state machine ensures the underlying prediction pipeline remains reliable and auditable. This paper situates the system within the current literature on BRFSS-based diabetes prediction, gradient boosting for tabular clinical data, explainable AI in healthcare, and LLM-driven health chatbots, and discusses the architectural decisions, evaluation results, limitations, and directions for future work.

Keywords: diabetes risk prediction; XGBoost; explainable artificial intelligence; SHAP; LIME; BRFSS; conversational AI; clinical decision support; machine learning; public health screening.

Table of Contents

1. Introduction

1.1 Background and Motivation

Diabetes mellitus, particularly Type 2 diabetes (T2DM), is a leading chronic condition worldwide, associated with substantial morbidity, mortality, and healthcare expenditure. A large fraction of individuals with T2DM or prediabetes remain undiagnosed for years, as clinical diagnosis typically relies on laboratory glucose or HbA1c testing that requires a healthcare visit. However, behavioural and demographic risk factors such as elevated body mass index (BMI), advancing age, hypertension, dyslipidaemia, physical inactivity, and self-reported general health are well-known correlates of diabetes risk and can be captured through simple, non-invasive questionnaires. This creates an opportunity for machine learning (ML) systems that translate widely available survey-style data into individualised risk estimates, enabling earlier referral for clinical confirmation.

Large-scale public health surveys such as the CDC's Behavioral Risk Factor Surveillance System (BRFSS) provide an ideal substrate for this task: BRFSS is an annual, telephone-administered survey of U.S. adults capturing self-reported health behaviours, chronic conditions, and socio-demographic attributes at national scale. Prior work has repeatedly demonstrated that supervised ML models trained on BRFSS data can distinguish diabetic from non-diabetic respondents with useful discriminative power, motivating continued development of BRFSS-based risk-screening tools.

1.2 Problem Statement

Despite promising predictive performance reported in the literature, three practical gaps limit the translation of BRFSS-based diabetes models into usable screening tools. First, many studies report model performance in isolation without an accompanying deployable system, leaving a gap between research prototype and usable application. Second, the class imbalance inherent in population-level diabetes prevalence, together with the black-box nature of high-performing ensemble models such as gradient boosting, undermines both statistical reliability (favouring accuracy over recall) and clinician/patient trust, a model that cannot explain itself is difficult to act upon. Third, existing tools rarely offer an accessible, conversational interface that lowers the barrier to completing a risk assessment, particularly for users unfamiliar with multi-field digital forms.

This paper addresses these gaps by presenting a complete, deployed system: a machine learning model (XGBoost) trained on the 2023 BRFSS release, wrapped in a dual-layer explainability framework (SHAP and LIME), and delivered through both a structured web dashboard and a guided conversational assistant, with supporting infrastructure for individual and batch screening, exportable clinical-style reports, and historical analytics.

1.3 Objectives

1. To curate a clinically and behaviourally meaningful feature subset from the 2023 BRFSS dataset that balances predictive value against the practical burden of a user-facing questionnaire.
2. To train and evaluate multiple classification algorithms (Logistic Regression, Decision Tree, Random Forest, XGBoost, LightGBM) for T2DM risk prediction, and select the best-performing model optimised for recall given the asymmetric cost of false negatives in a screening context.
3. To integrate SHAP and LIME to provide global (population-level) and local (per-patient) explanations of model predictions, supporting clinical interpretability and user trust.
4. To design and implement a production-style, full-stack platform API, database, dashboard, and reporting pipeline that operationalises the model for both individual and batch (CSV) risk screening.
5. To design a conversational assistant that lowers the interaction barrier to completing a risk assessment while preserving the reliability of a deterministic prediction pipeline.

1.4 Contributions

This work contributes: (i) a comparative evaluation of five classifiers (Logistic Regression, Decision Tree, Random Forest, XGBoost, LightGBM) on a fourteen-feature BRFSS 2023 subset chosen specifically for questionnaire brevity; (ii) a dual explainability layer combining SHAP's theoretically grounded global attributions with LIME's locally faithful, model-agnostic per-instance explanations; (iii) a full end-to-end reference architecture covering authentication, single and batch prediction, PDF/CSV/email export, and analytics that can serve as a template for similar clinical screening tools; and (iv) a hybrid conversational-plus-deterministic interaction design, in which a large language model handles open-ended health questions while a fixed client-side state machine drives the actual data-collection and prediction workflow, avoiding the reliability failures associated with relying on a language model to track multi-turn structured state.

1.5 Paper Organisation

Section 2 reviews related literature on BRFSS-based diabetes prediction, gradient boosting for tabular clinical data, explainable AI, and conversational health assistants. Section 3 describes the dataset, feature engineering, and modelling methodology. Section 4 details the system architecture and implementation. Section 5 reports evaluation results. Section 6 discusses findings, limitations, and ethical considerations, and Section 7 concludes with directions for future work.

2. Literature Review

This review synthesises four interrelated bodies of work that motivate the design choices in this study: (i) machine learning approaches to diabetes prediction using the BRFSS dataset; (ii) gradient boosting, and specifically XGBoost and LightGBM, for tabular clinical prediction tasks; (iii) explainable AI techniques principally SHAP and LIME as applied to healthcare models; and (iv) the emerging use of conversational, LLM-based agents for health risk assessment and patient engagement.

2.1 Machine Learning for Diabetes Prediction on BRFSS Data

The BRFSS dataset has become a standard benchmark for diabetes risk modelling owing to its scale and the breadth of behavioural, demographic, and health-status variables it captures. Nguyen and Zhang compared Decision Tree, K-Nearest Neighbours, and Logistic Regression classifiers on the BRFSS 2015 release, using twenty-one lifestyle and health-related features, and found meaningful differences in predictive performance across simple baseline models, underscoring the importance of model selection even before more complex ensembles are considered.

Working with a similar feature set, a comparative study using GA-XGBoost combined with a stacking ensemble reported that careful feature selection narrowing from the full BRFSS variable set to roughly twenty-five closely related predictors both simplified training and improved predictive accuracy and interpretability, a finding directly relevant to this study's decision to restrict the model to fourteen features chosen for clinical relevance and low respondent burden. Other work applying hybrid feature selection with SMOTE and Tomek-link resampling to the BRFSS 2015 release, evaluating logistic regression, random forest, gradient boosting, XGBoost, LightGBM, CatBoost, KNN, and SVM, found that ensemble combinations of boosting and instance-based methods achieved the strongest accuracy and ROC-AUC, reinforcing the choice of a boosting-family model as the predictive core of a BRFSS-based system.

More recent studies have moved toward newer BRFSS releases and explicit interpretability. Work using the 2021 BRFSS release investigated class-imbalance-aware algorithms and data augmentation strategies, confirming that elevated BMI, higher blood pressure, and advanced age are consistently associated with diabetes risk, a pattern this study's own SHAP analysis is expected to corroborate. An interactive Dash-based application built on BRFSS integrated SHAP and LIME for dual-level interpretability and selected a recall-optimised LightGBM model trained on an undersampled dataset for deployment, explicitly prioritising sensitivity over raw accuracy a design philosophy this study adopts as well, given that a missed at-risk case in a screening tool is costlier than a false alarm. Using the 2023 BRFSS release specifically, a state-level study of Tennessee adults compared seven algorithms (logistic regression, SVM, KNN, decision tree, random forest, gradient boosting, and XGBoost) with SHAP-based interpretation, and a nationwide study using the same era of data applied AI-driven analysis to identify the most influential behavioural and demographic predictors of diabetes risk across the full adult U.S. population. Together, these works establish both the continuity of the BRFSS-based prediction literature into its most recent survey years and the now-standard practice of pairing a boosting-family classifier with a post-hoc explainability layer.

2.2 Gradient Boosting and XGBoost for Tabular Clinical Data

Gradient-boosted decision tree ensembles have become the dominant approach for structured/tabular prediction tasks in healthcare, generally outperforming both classical linear models and, for datasets of this scale, deep neural architectures, while remaining considerably cheaper to train and tune. XGBoost, introduced by Chen and Guestrin, and LightGBM, introduced by Ke et al., are two widely adopted implementations. XGBoost uses a level-wise tree growth strategy and regularised boosting objective, while LightGBM employs a leaf-wise growth with gradient-based one-side sampling. Both are well-suited to BRFSS data, in which the majority of predictors are ordinal or categorical survey responses rather than continuous clinical measurements.

XGBoost's clinical applicability has been demonstrated across domains beyond diabetes: a SHAP-based XGBoost classifier for thyroid cancer recurrence achieved 97% accuracy and an AUC of 0.99, with SHAP dependence analysis identifying treatment response and risk stratification as the dominant predictive features, closely matching established clinical expectations. Comparative work on breast cancer classification evaluating XGBoost, CatBoost, and LightGBM found that XGBoost and LightGBM improved discriminative performance (AUC) while preserving recall relative to baseline, and that SHAP could be layered onto all three boosting frameworks without modification, evidencing their compatibility with standard post-hoc explainability tooling. These findings support this study's selection of XGBoost and LightGBM as the predictive core: they offer competitive discriminative performance on tabular clinical/survey data, together with first-class SHAP integration.

2.3 Explainable AI in Healthcare: SHAP and LIME

The adoption of high-performing but opaque ML models in healthcare has been consistently limited by a lack of interpretability, motivating the field of explainable AI (XAI). SHAP, introduced by Lundberg and Lee, unifies a family of additive feature-attribution methods under a game-theoretic (Shapley value) framework, providing theoretically grounded, consistent global and local feature importance scores; for tree ensembles specifically, the TreeExplainer variant computes exact Shapley values efficiently, making SHAP practical for production-scale gradient-boosted models such as XGBoost and LightGBM. LIME, introduced by Ribeiro, Singh, and Guestrin, instead constructs a locally faithful, interpretable surrogate model around any single prediction by perturbing the input and observing the resulting change in model output, making it model-agnostic and well suited to explaining individual patient-level predictions independent of the underlying model family.

A systematic literature review of LIME applications in medical imaging analysed fifty-two studies and concluded that LIME integration measurably improved transparency and clinician trust in ML-based diagnostic and prognostic systems, while also noting known limitations namely that local surrogate explanations can be somewhat unstable under input perturbation and may not always represent global model behaviour faithfully. This motivates the dual-explainability design adopted in this study: SHAP is used to communicate consistent, theoretically grounded global feature importance across the patient population, while LIME is used to generate an intuitive, easily communicated explanation of any single patient's prediction, mitigating the respective weaknesses of each method when used alone.

2.4 Conversational AI and LLM-Based Health Risk Assessment

A parallel and more recent strand of literature examines large language models (LLMs) as an interaction layer for health risk assessment and patient engagement, motivated by the observation that traditional structured-data ML pipelines, however accurate, still require a user to complete a rigid multi-field form. A review of LLMs in medical chatbots surveys applications spanning symptom checking, health information delivery, and mental health support, while explicitly cautioning that LLM deployment in healthcare carries risk-management obligations distinct from those of traditional discriminative models a caution reflected in this study's decision to keep the LLM confined to open-ended conversation and explanation, while the actual structured data collection and prediction call are handled deterministically outside the language model.

Directly relevant precedent for chatbot-driven risk assessment comes from a study that developed a generative-AI mobile application using fine-tuned LLaMA2, Flan-T5, and T0 models to perform no-code, conversational COVID-19 severity risk assessment, benchmarked against traditional classifiers including logistic regression, XGBoost, and random forest; the authors frame this as a shift from structured-data-only prediction toward streaming, question-and-answer-driven risk assessment, closely paralleling the conversational assessment flow implemented in this study's own assistant.

Complementary work on an LLM-based diagnostic chatbot combining retrieval-augmented generation with chain-of-thought prompting reported that explicit reasoning traces improved both diagnostic accuracy and user-facing transparency relative to opaque symptom-checker baselines, reinforcing this study's design choice to have the assistant explain, in plain language, which factors are driving a given risk estimate rather than returning a bare probability. Finally, a systematic review of GPT-driven conversational platforms in maternal and newborn health engagement a domain that, like diabetes screening, blends risk stratification with sustained behavioural engagement found strong six-month user-retention rates (62–78%) and no serious reported harms, supporting the broader premise that conversational interfaces can meaningfully improve engagement with risk-screening tools without introducing material safety concerns, provided appropriate guardrails are maintained.

2.5 Synthesis and Research Gap

Taken together, the literature establishes that (a) BRFSS is a proven substrate for diabetes risk modelling across multiple survey years; (b) gradient-boosting ensembles, and XGBoost/LightGBM in particular, offer strong performance on this class of tabular, largely categorical survey data; (c) SHAP and LIME are complementary, well-validated tools for making such models interpretable to both clinicians and patients; and (d) LLM-based conversational interfaces are an increasingly credible complement rather than a replacement for structured ML risk assessment. However, few published systems combine all four elements into a single, deployed, end-to-end platform with individual and batch screening, exportable reporting, and a conversational front-end that is architecturally decoupled from the underlying deterministic prediction pipeline. This paper addresses that gap.

3. Methodology

3.1 Data Source

The predictive model is trained on the Centers for Disease Control and Prevention's 2023 Behavioral Risk Factor Surveillance System (BRFSS) public-use dataset, an annual, cross-sectional telephone survey of non-institutionalised U.S. adults. The 2023 release used in this study comprises 429,086 respondent records. The BRFSS target label distinguishes respondents who report a prior diabetes diagnosis from those who do not, and the raw survey exposes several hundred candidate variables spanning demographics, chronic conditions, healthcare access, and health behaviours.

3.2 Feature Selection

From the full BRFSS variable set, fourteen features were selected for the production model (Table 1). The selection criteria prioritised: (i) established clinical or epidemiological association with diabetes risk, consistent with prior BRFSS-based studies reviewed in Section 2.1; (ii) availability without laboratory testing, so that the resulting screening tool remains fully self-reportable; and (iii) minimising respondent burden, since a shorter questionnaire improves completion rates in both the structured web form and the conversational assistant described in Section 4. This deliberately narrower feature set fourteen variables versus the twenty-to-twenty-five typically retained in prior work trades a small amount of potential discriminative signal for a substantially faster, more completable user-facing assessment, an explicit design trade-off consistent with the feature-reduction findings reported for GA-XGBoost stacking approaches.

Table 1. Feature set used for model training and the user-facing risk assessment.

Feature Code	Description	Category	Data Type
_BMI5	Body mass index (computed from self-reported height/weight)	Anthropometric	Continuous
_AGE80	13-level reported age group (18–80+)	Demographic	Ordinal
SEXVAR	Sex	Demographic	Binary
_IMPRACE	Imputed race / ethnicity	Demographic	Categorical
GENHLTH	Self-rated general health (excellent–poor)	Health status	Ordinal
PHYSHLTH	Days of poor physical health in past 30 days	Health status	Continuous
SMOKE100	Smoked ≥100 cigarettes in lifetime	Behavioural	Binary
_TOTINDA	Any physical activity in past 30 days	Behavioural	Binary

Feature Code	Description	Category	Data Type
EDUCA	Highest level of education completed	Socio-economic	Ordinal
INCOME3	Annual household income bracket	Socio-economic	Ordinal
_RFHYPE6	Told by a doctor of high blood pressure	Clinical history	Binary
_RFCHOL3	Told by a doctor of high cholesterol	Clinical history	Binary
CHCKDNY2	Told by a doctor of kidney disease	Clinical history	Binary
_MICHD	History of coronary heart disease or heart attack	Clinical history	Binary

3.3 Data Preprocessing

1. Missing/unknown/refused survey codes were treated as missing values and imputed or excluded according to feature type (mode imputation for categorical items, median imputation for continuous items such as BMI and PHYSHLTH).
2. Ordinal and categorical BRFSS codes were retained in their native encoding, leveraging XGBoost's native categorical-feature handling rather than one-hot encoding, which avoids unnecessary dimensionality expansion and preserves ordinal information where present.
3. The binary diabetes outcome label was derived from the BRFSS diabetes-diagnosis item, excluding gestational-diabetes-only responses to keep the target specific to Type 2 (and undifferentiated Type 1/Type 2) diabetes.
4. The dataset exhibits substantial class imbalance, consistent with population-level diabetes prevalence (roughly one in eight to one in nine U.S. adults); this imbalance was addressed at the model-training level (Section 3.4) rather than through synthetic oversampling, to avoid the distributional distortion risks associated with SMOTE-style resampling noted in prior BRFSS studies.

3.4 Model Selection and Training

Five classification algorithms were trained and evaluated: Logistic Regression (a linear baseline), Decision Tree (a single tree for interpretability benchmark), Random Forest (a bagging ensemble of trees), XGBoost (a gradient-boosted tree ensemble with level-wise growth), and LightGBM (a gradient-boosted tree ensemble with leaf-wise growth). The selection of XGBoost and LightGBM was motivated by the comparative evidence reviewed in Section 2.2: strong performance on tabular clinical/survey data, efficient handling of categorical features (via one-hot encoding), and first-class compatibility with SHAP’s

TreeExplainer for efficient Shapley-value computation. All models were trained with class weighting to counteract label imbalance, and the decision threshold was tuned on a held-out validation split to favour recall deliberately accepting a higher false-positive rate in exchange for fewer missed at-risk respondents, consistent with the recall-first design philosophy adopted in comparable BRFSS explainable-AI systems. The dataset was partitioned into training (80%), validation (10%), and held-out test (10%) splits. Hyperparameters (for tree depth, learning rate, number of estimators, etc.) were tuned on the validation split using grid search, and final performance is reported on the untouched test split (Section 5).

3.5 Explainability Layer

Two complementary explainability methods are integrated into the platform. SHAP's TreeExplainer computes exact Shapley values for the XGBoost model, providing (a) a global ranking of feature importance across the full patient population, surfaced in the platform's analytics dashboard, and (b) a per-prediction attribution of how much each of the fourteen features pushed an individual's predicted probability up or down, surfaced alongside every single-patient result. LIME is used as a secondary, model-agnostic local explanation, constructing a locally faithful linear surrogate around each prediction; because LIME and SHAP are computed independently and via different mechanisms, agreement between the two provides an informal robustness check on any single explanation, consistent with the dual-explainability rationale established in Section 2.3.

3.6 Evaluation Metrics

Given the class imbalance and the screening context, the primary evaluation metrics are recall (sensitivity) and ROC-AUC, supplemented by precision, F1-score, and overall accuracy for completeness. Recall is prioritised because, in a screening tool intended to trigger further clinical evaluation, the cost of a false negative (a missed at-risk individual) is substantially higher than the cost of a false positive (an unnecessary clinical follow-up).

4. System Architecture and Implementation

The platform is implemented as a full-stack, production-style application rather than a research notebook, reflecting this study's objective of demonstrating a deployable diabetes-screening tool. Table 2 summarises the seven architectural layers; each is elaborated below.

Table 2. System architecture layers.

Layer	Responsibility & Technology
Presentation	Next.js (React) dashboard: multi-step prediction form, batch CSV upload, history, analytics visualisations (Recharts), report downloads, and the embedded "Dida" conversational assistant.
Application / API	FastAPI backend exposing REST endpoints for authentication, single and batch prediction, prediction history, dashboard analytics, PDF/CSV/email export, and chat.
Machine Learning	Serialized XGBoost classifier and preprocessing pipeline, loaded at API startup; SHAP TreeExplainer and LIME explainer invoked per prediction for interpretability output.
Conversational AI	"Dida" assistant: an LLM (Llama-3.1-8B-Instruct, served via API) handles open-ended diabetes questions and result explanation, while a deterministic client-side state machine drives the fourteen-question data-collection flow and the prediction call itself.
Data & Persistence	Relational database (async SQLAlchemy) storing user accounts, individual predictions (with SHAP values and recommendations), and batch job metadata, scoped per authenticated user.
Identity & Access	OAuth 2.0 (Google) authentication via NextAuth, with JSON Web Tokens attached to authenticated API requests; anonymous users may still obtain a prediction but cannot persist, download, or email results.
Reporting	Server-side PDF generation (ReportLab) for single-patient and batch executive-summary reports; CSV export of batch results filtered by risk tier; SMTP-based email delivery of the single-patient PDF report.

4.1 Prediction Pipeline

Both the structured web form and the conversational assistant ultimately invoke the same backend prediction endpoint, ensuring a single source of truth for risk computation regardless of interaction modality. On receipt of the fourteen input features, the API applies the stored preprocessing pipeline, obtains a class prediction and probability from the XGBoost model, computes the corresponding risk tier (Low / Moderate / High, thresholded on predicted probability), generates SHAP attributions for the individual prediction, and for authenticated users persists the prediction, its explanation values, and a natural-language recommendation to the database, returning a unique prediction identifier used subsequently for PDF download or email delivery.

4.2 Batch Prediction and Analytics

Recognising that the tool may be used for population-level or cohort-level screening (for example, by a clinic reviewing a patient roster), the platform supports CSV-based batch prediction: an uploaded file is validated against the expected fourteen-column schema, scored row-by-row through the same XGBoost pipeline, and persisted as a batch job with aggregate statistics (risk-tier counts and percentages, mean/median predicted probability, and global SHAP feature importance across the batch). The analytics dashboard visualises this aggregate output, including risk distribution by age group and by BMI range, computed directly from the saved per-row predictions for that batch so that the displayed breakdowns remain consistent with the underlying data rather than being derived from job-level summary statistics alone.

4.3 Dida: A Hybrid Conversational Assistant

A central design contribution of this platform is its conversational assistant, Dida, which addresses the interaction-barrier problem discussed in Section 2.4 while explicitly avoiding the reliability risks of delegating structured, multi-turn state tracking to a language model. The assistant is architected in two decoupled layers:

- Open-ended conversation layer: An instruction-tuned LLM (Llama-3.1-8B-Instruct) answers general questions about diabetes, explains a user's prior result and its SHAP-derived drivers in plain language, and offers prevention guidance, constrained by a system prompt that explicitly forbids diagnosis and directs users toward professional consultation.
- Deterministic assessment layer: When a user elects to check their risk either via an explicit request or a contextual affirmative response to the assistant's offer control passes to a fixed, client-side state machine that presents the same fourteen questions used by the structured web form, one at a time, using selectable options rather than free-text entry wherever the underlying feature is categorical. This design choice follows directly from the LLM medical-chatbot literature's caution regarding risk management in language-model deployments: rather than asking the LLM to remember which of fourteen fields it has already collected across a multi-turn conversation a task at which smaller instruction-tuned models were found to be unreliable during development the assistant delegates state-tracking entirely to conventional application code, and invokes the same backend prediction endpoint used by the structured form once all fields are collected.

This hybrid design mirrors the architectural pattern identified in the reviewed LLM-based COVID-19 risk-assessment application, in which a conversational front-end is layered over a still-deterministic underlying assessment, while additionally separating the free-text conversational surface from the structured data-collection surface entirely a stricter separation motivated by empirically observed reliability limitations of small open-weight LLMs when asked to track long-running structured state directly.

4.4 Result Presentation, Export, and Delivery

Every prediction whether obtained via the structured form or Dida is presented with a visual risk gauge, the model's predicted probability, the single most influential SHAP feature, and a plain-language recommendation. Authenticated users may download a formatted PDF report of any individual prediction or have it emailed to an address of their choosing; batch job owners may additionally download CSV extracts filtered by risk tier (high, moderate, low, or all) and an executive-summary PDF aggregating batch-level findings. Export and email functionality are restricted to authenticated, owning users, consistent with standard access-control practice for personally identifiable health-adjacent data.

4.5 Security and Access Control

Authentication is performed via Google OAuth 2.0, issuing a bearer JSON Web Token used to authorise all subsequent API calls. Prediction, export, and email endpoints verify not only that a request is authenticated but that the requested resource (a specific prediction or batch job) is owned by the requesting user, preventing cross-user data access. Anonymous, unauthenticated use of the single-prediction endpoint is permitted to support ungated screening, but anonymous predictions are not persisted, downloadable, or emailable.

5. Results and Evaluation

5.1 Predictive Performance

On the held-out test partition, the XGBoost model achieves the performance summarised in Table 3. The recall-optimised decision threshold reflects the design priority established in Section 3.6: in a screening context, minimising missed at-risk individuals is weighted above minimising false alarms.

Table 3. Held-out test set performance.

Metric	Held-out Test Value	Interpretation
Recall (Sensitivity)	75%	Three in four true at-risk respondents are correctly flagged
ROC-AUC	83%	Strong overall discriminative ability between risk classes

XGBoost achieved the highest ROC-AUC (0.829) and F1-score (0.448), with a recall of 79.7%. LightGBM delivered nearly identical ROC-AUC (0.829) and slightly higher recall (79.9%), making both suitable for deployment. The Decision Tree achieved the highest recall (80.6%) but at the cost of lower precision (28.3%) and accuracy (68.0%), indicating a higher false-positive rate. Logistic Regression, while having the highest accuracy (72.9%), exhibited lower recall (76.4%) and ROC-AUC (0.822). Given the screening focus, XGBoost was selected as the primary production model due to its best overall balance of recall and ROC-AUC.

5.2 Comparison with Related BRFSS-Based Studies

XGBoost achieved the highest ROC-AUC (0.829) and F1-score (0.448), with a recall of 79.7%. LightGBM delivered nearly identical ROC-AUC (0.829) and slightly higher recall (79.9%), making both suitable for deployment. The Decision Tree achieved the highest recall (80.6%) but at the cost of lower precision (28.3%) and accuracy (68.0%), indicating a higher false-positive rate. Logistic Regression, while having the highest accuracy (72.9%), exhibited lower recall (76.4%) and ROC-AUC (0.822). Given the screening focus, XGBoost was selected as the primary production model due to its best overall balance of recall and ROC-AUC.

Table 4. Indicative comparison with related BRFSS-based diabetes prediction studies.

Study (BRFSS-based)	Primary Model	Reported Headline Metric
This study (BRFSS 2023, 14 features)	XGBoost	79% recall / 83% ROC-AUC
Ensemble feature-optimisation study (BRFSS 2015)	XGBoost + KNN ensemble	94.8% accuracy / 0.989 ROC-AUC
Two-stage explainable ML study (Rwanda BRFSS-style)	XGBoost + SMOTE	97.1% accuracy
Two-stage explainable ML study (baseline stage)	Gradient boosting	86.6% accuracy / 17.0% recall

5.3 Explainability Outcomes

Global SHAP analysis across the training population consistently ranks general health status (GENHLTH), BMI (_BMI5), age group (_AGE80), and hypertension status (_RFHYPE6) among the most influential predictors of elevated diabetes risk, a pattern that closely mirrors the risk-factor associations reported across the BRFSS literature reviewed in Section 2.1, lending face validity to the model's learned behaviour. At the individual-prediction level, SHAP and LIME explanations were qualitatively found to agree on the dominant one or two drivers of a given prediction in the large majority of cases inspected during development, with occasional divergence on secondary, lower-magnitude features consistent with the known local-instability characteristics of LIME's perturbation-based surrogate discussed in Section 2.3, and reinforcing the value of presenting both explanations rather than relying on either in isolation.

5.4 Usability of the Conversational Assistant

Qualitative evaluation of the Dida assistant during development confirmed the central architectural hypothesis of Section 4.3: when the fourteen-question assessment was driven directly by free-form LLM dialogue, the assistant intermittently failed to reliably present selectable options for categorical fields, lost track of previously collected answers over long conversations, and did not consistently trigger the final prediction call failure modes consistent with the general caution, noted in Section 2.4, regarding

LLM reliability for structured, safety-relevant tasks. Migrating state ownership to a deterministic client-side flow, while retaining the LLM strictly for open-ended dialogue and contextual triggering (recognising an affirmative reply to its own offer to begin the assessment), eliminated these failure modes in subsequent testing while preserving the natural-language entry point that motivated the assistant's inclusion in the first place.

6. Discussion

6.1 Strengths

- A deliberately narrow, respondent-friendly feature set (fourteen variables) that preserves the core, literature-validated risk factors while minimising the burden of both the structured form and the conversational assessment.
- A dual explainability layer (SHAP for consistent global/per-instance attribution, LIME as an independent, model-agnostic local check) that supports the interpretability goals emphasised throughout the reviewed XAI-in-healthcare literature.
- A recall-prioritised decision threshold appropriate to a screening rather than diagnostic use case, explicitly trading some accuracy for sensitivity to at-risk individuals.
- A hybrid conversational architecture that gains the engagement benefits reported for LLM-based health chatbots in the literature while structurally avoiding the reliability failures associated with delegating multi-turn structured state to a language model.
- A complete, deployable reference architecture authentication, persistence, batch processing, analytics, and multi-channel reporting that extends beyond the scope of most published BRFSS prediction studies, which typically stop at model evaluation.

6.2 Limitations

- BRFSS is a self-reported telephone survey and is subject to recall bias, social-desirability bias, and non-response bias; predictions inherit these upstream data-quality limitations regardless of downstream model quality.
- The fourteen-feature model, by design, omits laboratory-confirmed measures (fasting glucose, HbA1c) and several secondary BRFSS predictors used in higher-accuracy comparator studies; this is a deliberate usability trade-off but does represent a ceiling on achievable discriminative performance relative to feature-richer models.
- The conversational assistant's open-ended dialogue layer depends on a relatively small (8-billion-parameter) instruction-tuned LLM for cost and latency reasons; larger models may further improve response quality and reduce edge-case failures in free-text understanding, though the deterministic assessment layer is unaffected by this choice.

- As with all BRFSS-derived models, findings reflect the U.S. adult population captured by that survey and should not be assumed to generalise without re-validation to other countries, healthcare systems, or age groups outside the surveyed range.
- The system is explicitly positioned as a screening aid, not a diagnostic tool; this distinction is communicated in-product but ultimately depends on appropriate downstream clinical interpretation.

6.3 Ethical and Responsible-Use Considerations

Consistent with the responsible-AI perspectives raised in the reviewed medical-chatbot and health-engagement literature, the platform embeds several safeguards: the assistant is explicitly prompted never to provide a diagnosis and to consistently recommend professional consultation; access to persisted predictions, PDF export, and email delivery is restricted to the authenticated owner of that data; and every prediction is presented alongside its explanation rather than as an unqualified numerical output, in line with the trust-building rationale for explainability established in Section 2.3. Future deployments handling real patient populations would additionally require formal clinical validation, regulatory consideration (depending on jurisdiction and intended use claims), and a documented data-governance and retention policy beyond the scope of this study.

7. Conclusion and Future Work

This paper presented the design, implementation, and evaluation of an explainable, full-stack Type 2 diabetes risk-screening platform built on the CDC's 2023 BRFSS survey. A XGBoost classifier trained on fourteen carefully selected features achieves 79% recall and 83% ROC-AUC, a deliberately recall-favouring result appropriate to a screening context. The platform integrates SHAP and LIME to render both global and per-patient predictions interpretable, and extends beyond model evaluation into a deployed reference architecture covering authentication, individual and batch prediction, analytics, and multi-channel (PDF/CSV/email) reporting. A key contribution is the hybrid design of the platform's conversational assistant, Dida, which combines an LLM-driven open-ended dialogue layer with a deterministic, client-side data-collection state machine an architecture motivated by, and consistent with, documented reliability concerns around LLM-driven structured health assessment in the wider literature.

Future work includes: (i) incorporating laboratory-derived features where available to assess the marginal predictive gain against the current self-report-only feature set; (ii) formal prospective or retrospective clinical validation against confirmed diagnoses rather than self-reported outcomes; (iii) extending the explainability layer with counterfactual explanations ("what change in BMI or activity level would move this prediction to a lower risk tier") to make recommendations more directly actionable; (iv) evaluating larger or fine-tuned LLMs for the conversational layer to assess whether response quality can

be improved without reintroducing the structured-state reliability issues that motivated the current architecture; and (v) conducting a formal, quantitative usability study of the conversational assessment flow against the traditional structured form, extending the qualitative findings reported in Section 5.4.

References

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. Electronics, 8(8), 832.

International Diabetes Federation. (2021). IDF Diabetes Atlas (10th ed.). https://www.diabetesatlas.org

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.

Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future — big data, machine learning, and clinical medicine. New England Journal of Medicine, 375(13), 1216–1219.

Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. New England Journal of Medicine, 380(14), 1347–1358.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144.