

Paper Type: Original Article

An AI-Enabled Integrated Framework for Diabetes Risk Screening, Self-Management, and Care Navigation

EL-Sayed Ramy ¹ , Esraa Mohsen ¹ , Rana Sameh ¹ , Romisah Osama ¹ , Farouk Mohamed ¹ ,
Youssef Emad ¹  and Walid Abdullah ^{1,*} 

¹ Faculty of Computers and Informatics, Zagazig University, Zagazig 44519, Sharqiyah, Egypt;

Emails: Sayedramy45@gmail.com, esraamohsen042@gmail.com, ranasameh290@gmail.com, romamac25@gmail.com, faroukmohamed1011@gmail.com, Youssefemad9336@gmail.com, waleed@zu.edu.eg.

Received: 01 May 2026

Revised: 16 Jul 2026

Accepted: 03 Oct 2026

Published: 07 Oct 2026

Abstract

Diabetes management applications commonly emphasize logging and reminders, whereas machine-learning studies often evaluate prediction models outside the workflow in which users act on their outputs. This paper presents Diabetes, a mobile platform that integrates tabular risk screening, self-management support, Arabic-language educational assistance, and location-aware care navigation within one deployable architecture. The system combines a Flutter client with Node.js/Express services, hybrid Firebase/PostgreSQL persistence, three task-specific classifiers, and OpenStreetMap/OSRM-based doctor routing. The predictive layer is evaluated on three public tabular benchmarks. On the archived internal test splits, the Random Forest general-risk model achieved 94.10% accuracy, 90.11% sensitivity, 98.10% specificity, and ROC-AUC 0.9916; the SVC women-focused risk model achieved 90.29% accuracy and ROC-AUC 0.95; and the XGBoost health-indicator model achieved 94.99% accuracy, 95.00% sensitivity, 94.98% specificity, and ROC-AUC 0.9900. The publication version deliberately distinguishes risk screening from clinical diagnosis. In particular, the 1,000-record women-focused benchmark contains pregnancy-count and metabolic variables but does not constitute a gestational-diabetes cohort; it is therefore not presented as evidence of gestational-diabetes diagnostic validity. The study's contribution is an end-to-end digital-health architecture that couples model outputs to monitoring, educational support, and clinician navigation while making evaluation boundaries explicit. Because the reported results arise from public benchmark data and internal development splits, they should be interpreted as proof-of-concept performance rather than prospective clinical validation.

Keywords: Diabetes Risk Screening; Mobile Health; Machine Learning; Random Forest; Support Vector Machine; XGBoost; Self-Management; Arabic Digital Health; Clinical Decision Support.

1 | Introduction

Diabetes remains a major global health burden. The 11th edition of the International Diabetes Federation (IDF) Diabetes Atlas estimates that 589 million adults aged 20–79 years were living with diabetes in 2024 and that approximately 252 million of them were unaware of their condition [1,2]. The burden is especially relevant to the Middle East and North Africa, which has one of the highest age-standardized regional prevalences reported by the IDF [2]. These figures strengthen the case for accessible prevention, timely risk assessment,



Corresponding Author: waleed@zu.edu.eg



<https://doi.org/10.61356/j.scin.2026.3675>



Licensee SciNexuses. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0>).

and sustained self-management support, but they do not change the clinical requirement that diabetes diagnosis be based on validated diagnostic criteria and professional interpretation [3].

Mobile health is a practical channel for supporting daily behavior because diabetes management depends on repeated actions rather than a single clinical encounter. Recent meta-analytic evidence indicates that diabetes self-management applications can contribute to glycemic improvement when they support behaviors such as self-monitoring and medication-taking [4], while a 2025 systematic review found continuing interest in mobile delivery of diabetes self-management education and support [5]. Digital tools are therefore most useful when they combine low-friction logging, reminders, understandable feedback, and a clear path to professional care.

At the same time, the machine-learning literature on diabetes prediction contains a recurrent translation gap. Models are frequently assessed on public benchmark datasets, yet deployment papers may move too quickly from internal discrimination metrics to claims of diagnosis or clinical reliability. Contemporary reporting guidance such as TRIPOD+AI emphasizes transparent description of participants, predictors, outcomes, analysis, and model evaluation [6], while PROBAST+AI explicitly separates the quality of model development from the risk of bias in performance evaluation [7]. Those distinctions are particularly important when class rebalancing, synthetic oversampling, feature selection, or public benchmark data are used.

The system studied here was developed as an integrated mobile application rather than as a stand-alone classifier. It combines three task-specific risk models with medication and lifestyle reminders, an Arabic-capable conversational assistant, and location-aware doctor discovery. The publication contribution is therefore twofold: it reports the predictive outputs that are actually available in the project artifacts, and it places those outputs inside a safety-conscious digital-health workflow that does not equate an algorithmic score with a medical diagnosis. This framing is consistent with WHO guidance that AI for health should preserve human oversight, transparency, accountability, and appropriate limits on automated decisions [8].

1.1 | Research Gap and Scientific Contribution

The research gap addressed by this work is the operational fragmentation between risk modeling and self-management. Prediction-only studies can achieve strong internal metrics but do not necessarily show how a user records daily behavior, interprets a risk estimate, locates a clinician, or receives accessible educational support. Conversely, many self-management applications concentrate on logging, visualization, and reminders without exposing how predictive signals are generated or evaluated. The proposed platform connects these functions through a modular architecture while preserving task boundaries.

Table 1. Scientific positioning and contributions of the proposed platform.

Dimension	Common treatment in prior work	Risk or gap	Contribution in this study
Prediction models	Single tabular model evaluated on one dataset.	Strong internal metrics can be mistaken for clinical validity.	Three task-specific models are reported with confusion-matrix-derived metrics and an explicit proof-of-concept scope.
Self-management	Logging, reminders, education, or device integration [4,5,9].	Behavioral support may be disconnected from risk stratification.	Risk outputs, reminders, activity/meal logging, and educational support are connected in the same mobile workflow.
Women-focused risk	Pregnancy-related variables are sometimes treated as proxies for gestational diabetes.	A pregnancy-count feature does not establish a gestational-diabetes cohort.	The second benchmark is correctly framed as women-focused diabetes risk, not validated GDM diagnosis.

Care navigation	Provider discovery is often outside the prediction workflow.	Users may receive a risk signal without a concrete route to care.	Haversine pre-filtering and OSRM road routing support location-aware specialist discovery.
Language and governance	Digital-health interfaces may not match local language needs.	Poor accessibility and overconfident health messaging increase deployment risk.	Arabic support is combined with a non-diagnostic decision-support boundary and clinician referral pathway.

As summarized in Table 1, the novelty claimed here is architectural and workflow-oriented rather than a claim that Random Forest, SVC, or XGBoost are new algorithms. The scientific value lies in the disciplined separation of three risk-estimation tasks, the integration of those tasks with daily self-management and care navigation, and the explicit correction of clinical overclaims that would otherwise weaken reproducibility and safety.

1.2 | Research Questions

The evaluation is organized around four questions. RQ1 asks how accurately the general diabetes model separates positive and negative cases on the archived internal test split. RQ2 asks whether a second women-focused classifier provides consistent discrimination on a smaller benchmark. RQ3 asks whether a health-indicator model based on survey-derived lifestyle and comorbidity variables can maintain balanced sensitivity and specificity after project preprocessing. RQ4 asks how these predictive components can be integrated with reminders, education, and location-aware care navigation without representing benchmark predictions as definitive medical diagnoses.

2 | Related Work

2.1 | Mobile Diabetes Self-Management

Evidence for mobile diabetes support increasingly focuses on the behavior-change mechanisms embedded in applications rather than on the mere presence of a mobile interface. Tarricone et al. synthesized randomized trials of diabetes self-management apps and reported a pooled reduction in HbA1c favoring app-based interventions, with self-monitoring and medication-related behavior among the components associated with improved outcomes [4]. Versluis et al. likewise reviewed mobile delivery of diabetes self-management education and support, reinforcing the importance of structured content and ongoing engagement [5]. These findings support the inclusion of reminders, logging, feedback, and understandable educational material in the present system.

A concrete example of a mature digital logbook is mySugr. A multicenter randomized trial evaluated a digital diabetes logbook in people with type 1, type 2, and gestational diabetes [9], and a subsequent real-world study in India examined glycemic outcomes associated with mySugr use [10]. Such studies demonstrate the clinical value of structured digital self-management, but they also illustrate a distinction central to this paper: a self-management application with clinical evidence is not the same as a prediction model trained on a public benchmark, and the evidentiary standard for each function should remain visible.

2.2 | Machine Learning for Diabetes Risk

The predictive methods used in this project are established machine-learning approaches. Random Forest combines multiple decision trees and reduces variance through ensemble aggregation [11]. Support vector machines construct a separating hyperplane with maximum margin and remain effective for medium-sized tabular datasets when scaling and hyperparameter selection are handled carefully [12]. XGBoost implements regularized gradient tree boosting and is widely used for structured data because it can capture nonlinear

interactions while controlling model complexity [13]. Class imbalance can be addressed with resampling methods such as SMOTE, which synthetically generates minority-class observations [14], although oversampling must be restricted to training data to prevent optimistic evaluation.

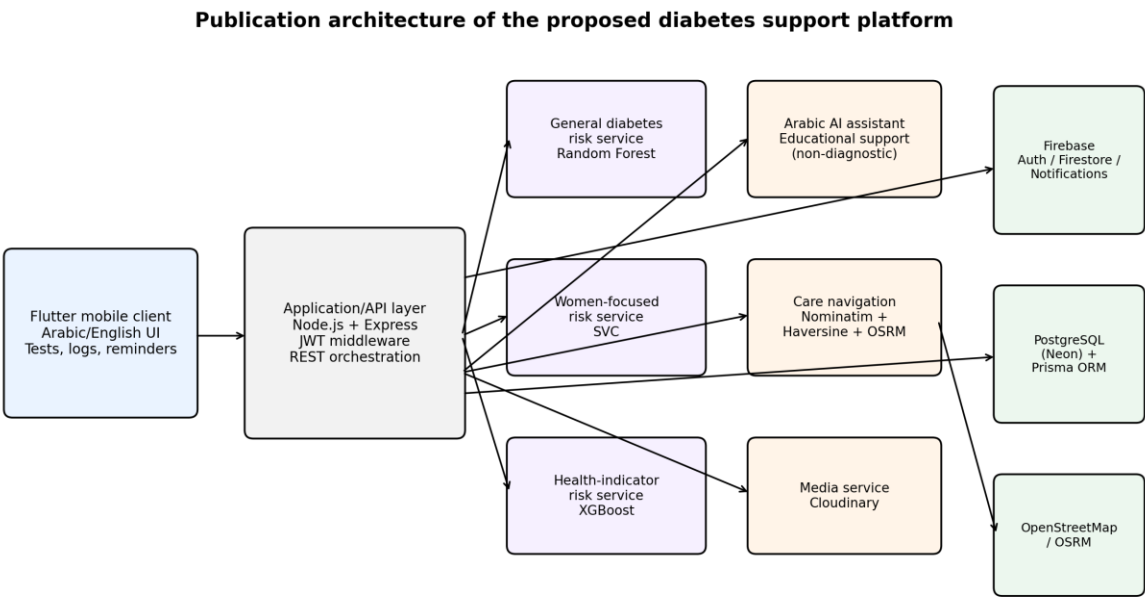
For clinical prediction, however, algorithm choice is only one part of methodological quality. TRIPOD+AI calls for clear reporting of data sources, outcome definitions, handling of missing data, sample size, model specification, evaluation, and limitations [6]. PROBAST+AI further highlights participant/data-source quality, predictors, outcome definition, and analysis as core domains of bias and applicability [7]. These principles motivate the present paper’s conservative interpretation of benchmark performance and its refusal to relabel a general diabetes dataset as a clinically validated gestational-diabetes cohort.

2.3 | Positioning of the Proposed System

The proposed application occupies the intersection of three streams: predictive risk estimation, diabetes self-management, and care navigation. The system does not attempt to replace laboratory testing, obstetric screening protocols, or professional diagnosis. Instead, predictive models are used as front-end risk signals that can trigger educational messages or encourage a clinician visit. This distinction is especially important because current clinical guidance defines diabetes and gestational-diabetes screening using specific laboratory criteria and pregnancy-timing considerations that are not represented in all public machine-learning datasets [3].

3 | System Design and Software Architecture

Figure 1 presents the publication architecture reconstructed from the implemented components described in the project. The mobile client is developed in Flutter. Requests are coordinated through Node.js/Express REST services, with PostgreSQL/Neon and Prisma used for structured relational persistence and Firebase services supporting authentication, real-time/mobile data functions, and notifications. Cloudinary is used for media storage. The predictive services are separated from the application layer so that model replacement does not require redesign of the entire mobile client. The same service boundary also isolates the conversational assistant and the doctor-search pipeline.



Predictive outputs are framed as risk-screening signals and require clinical confirmation for diagnosis.

Figure 1. Publication architecture of the proposed platform. Predictive services, self-management functions, conversational support, and location-aware care navigation are separated behind the application/API layer.

3.1 | Functional Modules

The platform contains four operational groups. The first is risk screening: users provide the variables required by the selected model and receive a risk-oriented result. The second is self-management: medication, meal, tests, and activity reminders support recurring daily tasks. The third is educational assistance: the Arabic-capable conversational interface provides diabetes-related explanations and general guidance. The fourth is care navigation: users can search for a doctor's directory, inspect provider information, and identify nearby options. These functions are intentionally presented as a continuum from awareness to professional follow-up rather than as autonomous diagnosis or treatment.

3.2 | Location-Aware Doctor Search

The doctor-search workflow combines fast geographic filtering with road-network routing. Candidate locations are geocoded using Nominatim. A Haversine-distance calculation provides an inexpensive first-stage proximity estimate, after which the nearest candidates can be refined using OSRM road-network distances. OSRM was designed for real-time routing over OpenStreetMap data and supports efficient shortest-path computation at large scale [15]. The Haversine distance between two latitude-longitude pairs is computed as:

$$d = 2R \arcsin(\sqrt{\sin^2((\phi_2 - \phi_1)/2) + \cos(\phi_1) \cos(\phi_2) \sin^2((\lambda_2 - \lambda_1)/2)})$$

where R is Earth's mean radius, ϕ denotes latitude, and λ denotes longitude. The Haversine stage is used only as a spatial pre-filter; road distance is more relevant for final navigation because physical travel follows the road network rather than a great-circle path.

3.3 | Safety Boundary and Intended Use

The intended use is preliminary risk support and self-management, not autonomous clinical diagnosis. A high-risk result should be interpreted as a reason to seek clinical assessment, while a low-risk result does not rule out diabetes. The conversational assistant is similarly positioned as an educational interface, not as a prescriber or emergency service. This scope is aligned with WHO guidance that health AI should preserve autonomy, transparency, accountability, and human oversight [8].

4 | Materials and Methods

4.1 | Data Sources and Provenance

Three public tabular datasets were used in the original project. Their roles are separated in Table 2 because they represent different populations, variables, and outcome constructions. Treating these datasets as interchangeable would make the evaluation clinically ambiguous. The publication version therefore reports each model only against the dataset on which its archived output was generated.

Table 2. Dataset provenance, task definition, and publication scope.

Task	Public source	Source size	Predictor profile	Archived evaluation	Publication scope
General diabetes risk	Kaggle diabetes prediction dataset [16]	100,000 rows; 8 candidate predictors + target	Age, sex/gender, BMI, hypertension, heart disease, smoking history, HbA1c, blood glucose	16,716-case balanced internal test output	General diabetes risk screening; not prospective diagnosis.

Women-focused diabetes risk	Kaggle “Diabetes Prediction” benchmark [17]	1,000 rows; 8 predictors + target	Pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, pedigree function, age	278-case balanced internal test output	Women-focused diabetes classification. Not a validated GDM cohort.
Health-indicator risk	BRFSS 2015 derivative on Kaggle [18]	70,692-row balanced binary file; 21 predictors + target	Blood pressure, cholesterol, BMI, smoking, activity, general health, sociodemographic indicators, etc.	6,213-case internal test output after project filtering/feature selection	Survey-based diabetes/prediabetes risk signal; not a daily physiological score.

Table 2 also resolves two terminology problems in the original manuscript. First, the 1,000-record dataset contains pregnancy count but no gestational-age variable, current-pregnancy indicator, or gestational-diabetes outcome; it therefore cannot support a claim of validated GDM diagnosis. Second, the BRFSS model predicts a diabetes-related class from health indicators; it is more accurately described as a health-indicator risk classifier than as a continuously measured “daily health score.” These corrections make the scientific interpretation consistent with the available data.

4.1.1 | General Diabetes Benchmark

The first dataset is the public diabetes prediction benchmark containing demographic, comorbidity, and metabolic variables [16]. The archived project pipeline applied cleaning, encoding/scaling, outlier handling, and class balancing. The final model-stage output recorded 66,861 negative and 5,772 positive observations before SMOTE and a balanced test output of 8,358 cases per class. Because the archived artifact does not preserve a fully executable data-splitting ledger, the publication reports the resulting test confusion matrix exactly but treats it as an internal development estimate rather than an external validation result.

4.1.2 | Women-Focused Diabetes Benchmark

The second dataset contains 1,000 records with the variables Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age, and a binary diagnosis label [17]. These variables resemble Pima-style diabetes benchmarks and support a women-focused diabetes classification task. They do not define gestational diabetes. The SVC results are therefore preserved as a secondary diabetes-risk experiment, while the pregnancy-specific diagnostic claim is removed from the publication version.

4.1.3 | BRFSS Health-Indicator Benchmark

The third dataset is derived from the 2015 Behavioral Risk Factor Surveillance System (BRFSS) and is distributed as a cleaned diabetes health-indicator dataset [18]. The public source includes a 70,692-row balanced binary subset with 21 predictors and a target. In the project pipeline, duplicate/anomaly filtering reduced the working data to 43,702 observations before final modeling, and feature selection retained 12 variables: HighBP, GenHlth, HighChol, Age, BMI, DiffWalk, Income, Education, HeartDiseaseorAttack, PhysActivity, PhysHlth, and Smoker. The model output is interpreted as survey-based risk discrimination rather than a real-time physiological monitor.

4.2 | Model Development

4.2.1 | Random Forest General-Risk Model

The general-risk model uses a Random Forest classifier with 100 trees and otherwise default project settings. Random Forest was selected because ensemble averaging can reduce the instability of a single decision tree and represent nonlinear interactions among demographic and metabolic predictors [11]. The archived project report does not retain numerical outputs for every exploratory baseline, so the present paper does not invent

comparative baseline values; it reports only the final model for which a confusion matrix and ROC curve are available.

4.2.2 | Support Vector Classifier for Women-Focused Risk

The second model uses an SVC with hyperparameters selected by GridSearchCV. The archived search space includes C in $\{0.1, 1, 10\}$, kernels $\{rbf, poly\}$, and gamma in $\{scale, auto, 0.1, 1\}$, with five-fold cross-validation and ROC-AUC as the optimization objective. Feature scaling is appropriate for margin-based methods because predictor scale affects the geometry of the decision boundary [12]. The output is reported as women-focused diabetes risk rather than gestational-diabetes diagnosis.

4.2.3 | XGBoost Health-Indicator Model

The health-indicator model uses XGBoost after project preprocessing and feature selection. The workflow described in the development artifacts includes duplicate handling, anomaly detection, scaling, class rebalancing, and SelectKBest/ANOVA-based feature selection before final fitting. XGBoost's regularized boosting objective is well suited to heterogeneous tabular predictors and can model nonlinear interactions without requiring a manually specified parametric form [13].

4.2.4 | Resampling and Leakage Control

SMOTE and related resampling techniques can be useful for development on imbalanced data [14], but they can also create optimistic performance if synthetic observations are generated before the train/test boundary. The archived project outputs show balanced test sets for the first two models, and the exact order of resampling versus splitting is not recoverable from the manuscript alone. Consequently, the reported metrics are treated as conditional internal estimates. A clinical-grade re-evaluation would require splitting first, fitting all preprocessing and resampling operations only on training folds, and then testing on untouched observations with the natural target prevalence. This is a methodological boundary of the available evidence, not a reason to fabricate a cleaner experimental history.

4.3 | Evaluation Metrics

Performance is reconstructed from the archived confusion matrices. Let TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives. Accuracy, sensitivity (recall), specificity, precision, and F1-score are defined as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (2)$$

$$Specificity = \frac{TN}{TN + FP} \quad (3)$$

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

$$F1 = 2 * (Precision * Sensitivity) / (Precision + Sensitivity) \quad (5)$$

ROC-AUC is taken from the archived ROC plots. For accuracy, a nominal 95% Wilson interval is computed from the reported test-set counts. These intervals quantify binomial uncertainty conditional on the recorded test split; they do not correct for possible resampling leakage, dataset shift, prevalence shift, or external-validation uncertainty.

4.4 | Implementation Environment

The mobile application is implemented in Flutter with Provider-based state management. The service layer uses Node.js and Express with REST endpoints, JWT-oriented middleware, and Python-based model services. Structured data are stored using PostgreSQL/Neon with Prisma ORM, while Firebase services support authentication, real-time/mobile functions, and notifications. Cloudinary is used for image/media storage, and Docker/Fly.io support containerized deployment. Doctor discovery combines OpenStreetMap/Nominatim data with OSRM routing. This modular organization allows the prediction services to be revised independently of the user interface and persistence layers.

5 | Results

5.1 | General Diabetes Risk Model

Figure 2 shows the archived confusion matrix and ROC curve for the Random Forest model. The test output contains 8,358 negative and 8,358 positive cases. Of the negative cases, 8,199 were classified correctly and 159 were false positives. Of the positive cases, 7,531 were true positives and 827 were false negatives. These counts yield 94.10% accuracy, 90.11% sensitivity, 98.10% specificity, 97.93% positive predictive value on the balanced test set, and an F1-score of 93.86%. The archived ROC-AUC is 0.9916. The gap between sensitivity and specificity is clinically relevant: within this development split, the model misses more positive cases than it incorrectly flags negative cases.

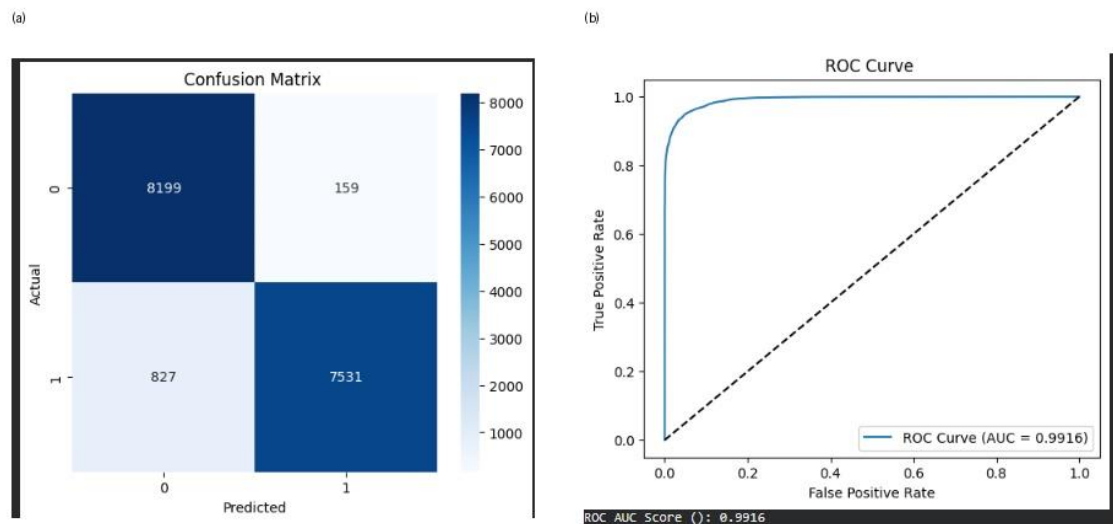


Figure 2. General diabetes risk model: (a) confusion matrix and (b) archived ROC curve (AUC = 0.9916).

5.2 | Women-Focused Diabetes Risk Model

The SVC evaluation contains 278 cases, balanced at 139 per class. As Figure 3 shows, 129 negative cases were correctly classified and 10 were false positives; 122 positive cases were correctly classified and 17 were false negatives. The corresponding accuracy is 90.29%, sensitivity 87.77%, specificity 92.81%, precision 92.42%, and F1-score 90.04%. The archived ROC-AUC is 0.95. The smaller sample size produces visibly wider uncertainty than the other experiments, and the dataset does not justify a gestational-diabetes interpretation.

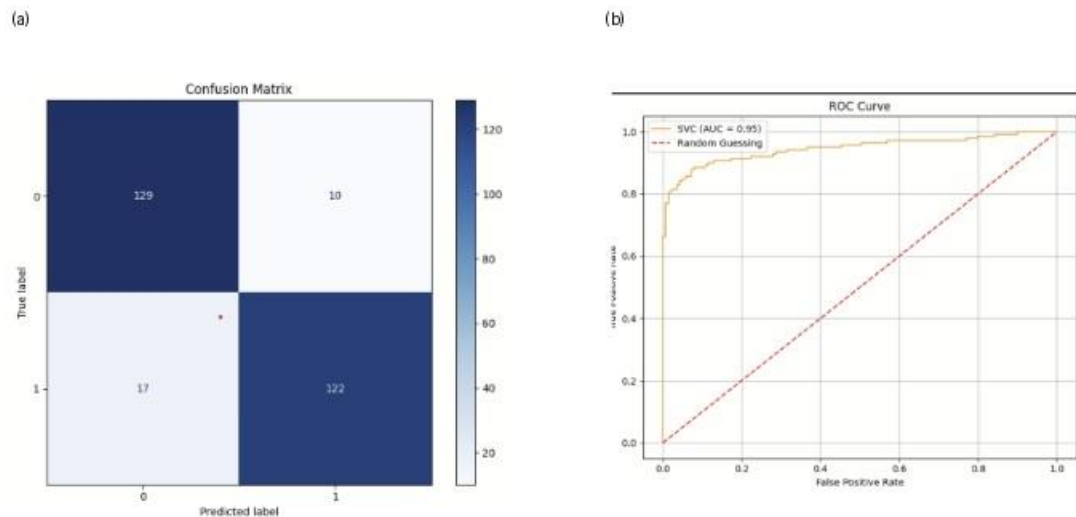


Figure 3. Women-focused diabetes risk model: (a) confusion matrix and (b) archived ROC curve (AUC = 0.95). The dataset is not treated as a gestational-diabetes cohort.

5.3 | Health-Indicator Risk Model

Figure 4 presents the XGBoost results after the project's anomaly filtering and 12-feature selection. Of 3,090 negative cases, 2,935 were true negatives and 155 were false positives. Of 3,123 positive cases, 2,967 were true positives and 156 were false negatives. This produces 94.99% accuracy, 95.00% sensitivity, 94.98% specificity, 95.04% precision, and an F1-score of 95.02%. The archived ROC-AUC is 0.9900. Unlike the first model, sensitivity and specificity are almost identical on the reported test split.

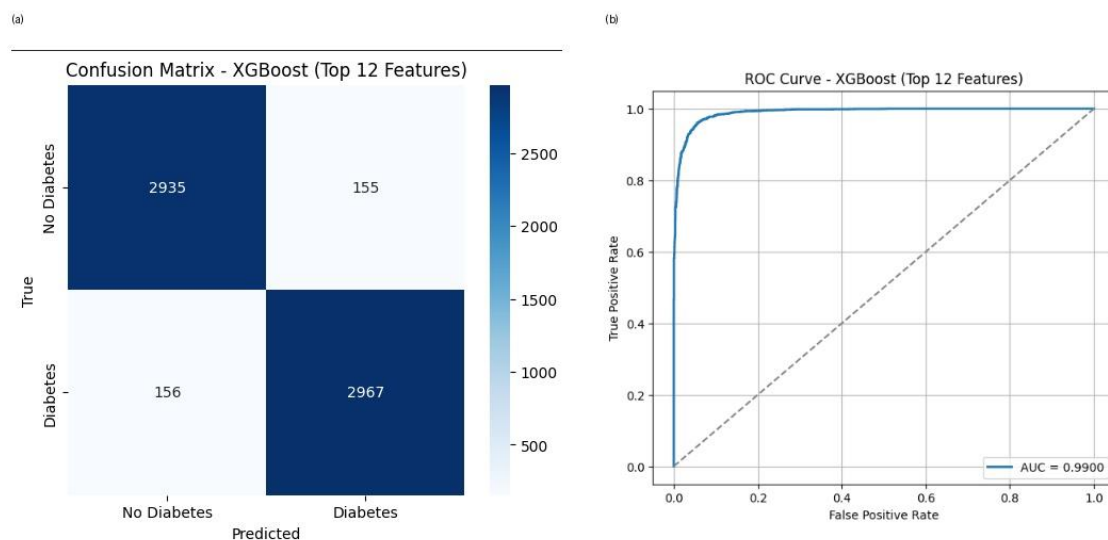


Figure 4. Health-indicator risk model: (a) confusion matrix and (b) archived ROC curve for the XGBoost model (AUC = 0.9900).

5.4 | Consolidated Quantitative Results

Table 3 permits exact comparison of the recorded internal outputs but should not be read as a ranking of clinical usefulness. The models were evaluated on different datasets with different feature spaces and artificially balanced or filtered test distributions. The strongest defensible conclusion is that all three

development pipelines show high discrimination on their respective archived internal splits; external performance remains unknown.

Table 3. Consolidated internal test performance reconstructed from archived confusion matrices.

Model	n	Accuracy (95% CI)	Sensitivity	Specificity	Precision	F1	ROC-AUC
General risk / Random Forest	16716	94.10% (93.73–94.45)	90.11%	98.10%	97.93%	93.86%	0.9916
Women-focused risk / SVC	278	90.29% (86.24–93.24)	87.77%	92.81%	92.42%	90.04%	0.9500
Health-indicator risk / XGBoost	6213	94.99% (94.42–95.51)	95.00%	94.98%	95.04%	95.02%	0.9900

5.5 | Integrated Application Workflow

The predictive functions are embedded in a broader mobile workflow. Figure 5 illustrates the implemented user interface for home access, doctor discovery, physician profile inspection, Arabic conversational support, reminder configuration, and risk-result/educational screens. The workflow reduces the number of separate tools a user needs for risk awareness, routine organization, and provider discovery. Importantly, the user-facing result should be interpreted as a screening signal followed by education or referral, not as a replacement for laboratory confirmation.

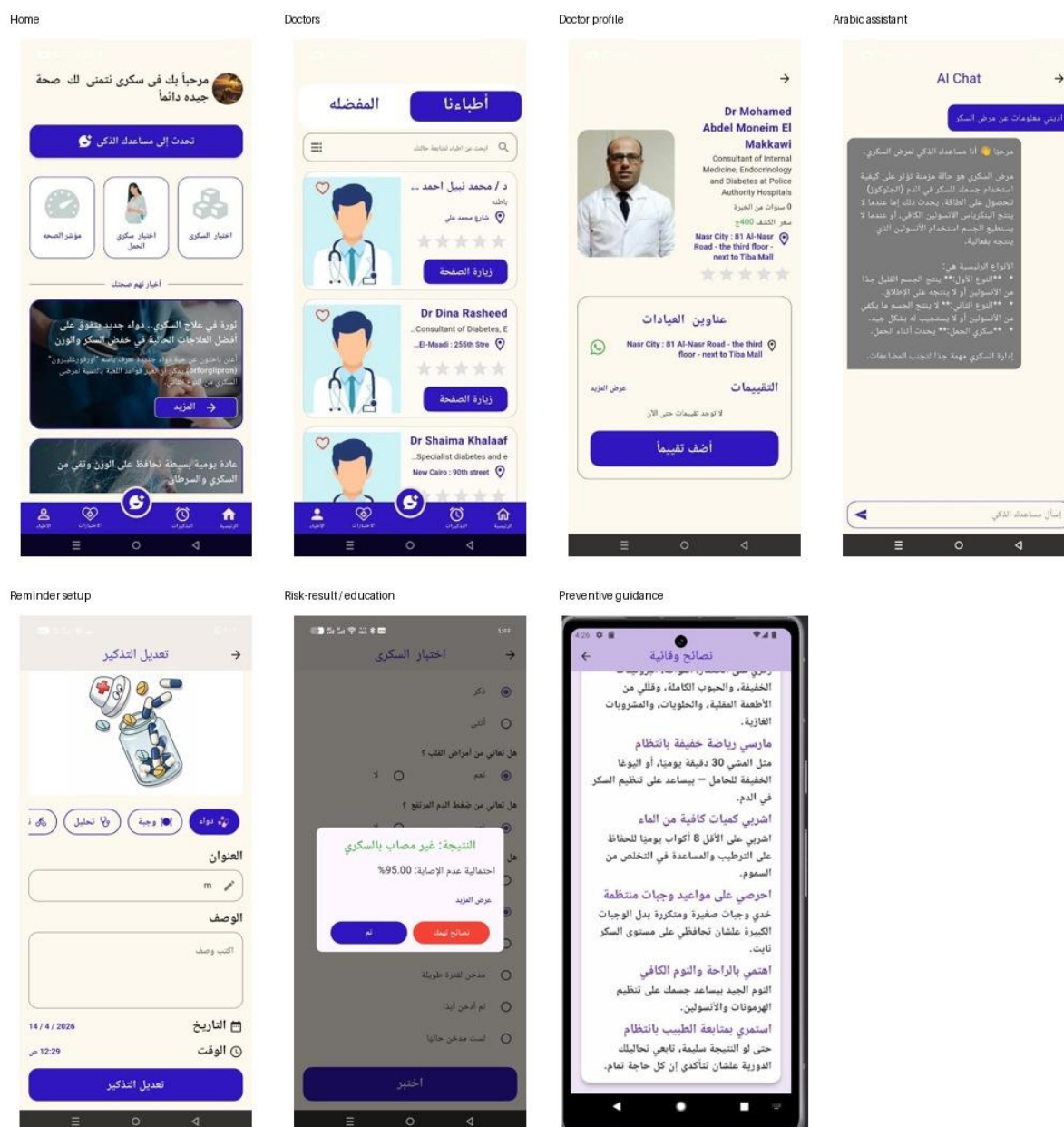


Figure 5. Representative mobile interfaces from the implemented system: home access, doctor discovery and profile, Arabic conversational support, reminders, risk-result presentation, and preventive education.

6 | Discussion

6.1 | Interpretation of the Predictive Results

The general Random Forest and the health-indicator XGBoost model both show high ROC-AUC values on the archived internal splits. Their error profiles, however, differ. The general model is more specific than sensitive, whereas the health-indicator model is nearly symmetric across the two classes. For screening-oriented applications, sensitivity deserves particular attention because a false negative can incorrectly reassure a user who may need further assessment. A deployment threshold should therefore be selected according to clinical cost and prevalence rather than by maximizing accuracy alone.

The SVC experiment is methodologically useful because it demonstrates performance on a much smaller tabular dataset, but it also illustrates why variable names must not be overinterpreted. A “Pregnancies” predictor records pregnancy history; it does not establish that the sample consists of currently pregnant people

or that the target is gestational diabetes. The publication version therefore preserves the numerical experiment while removing a clinical label that the data cannot support. This correction materially strengthens the paper because it aligns the stated task with the observed predictors and outcome.

The health-indicator model should likewise be understood as a risk classifier derived from survey-style attributes. It is not a sensor-based continuous health index. Its value lies in translating a compact set of health and lifestyle indicators into a binary risk signal that can be coupled to educational guidance and care navigation. The selected features—blood pressure, cholesterol, BMI, general health, mobility, age, income, education, heart-disease history, activity, physical health, and smoking—span both biomedical and social dimensions, which is consistent with the multifactorial nature of diabetes risk.

6.2 | Why Integration Matters

The platform-level contribution is stronger than any individual classifier. Evidence from diabetes mHealth trials suggests that recurring self-management behaviors, not one-time prediction alone, drive practical benefit [4,5]. The application therefore links a risk signal to reminders, daily logs, educational content, and doctor navigation. This creates an actionable sequence: risk awareness can motivate monitoring; monitoring can reveal persistent concerns; educational support can improve understanding; and the care-navigation module can reduce friction in reaching professional evaluation.

The architecture also separates concerns. The mobile client does not embed model logic directly; instead, it calls isolated services through the backend. This permits a future clinically validated model to replace a benchmark model without rewriting the application, and it makes versioning, monitoring, and rollback more feasible. The same separation is important for the conversational assistant, which should be governable independently from the prediction layer.

6.3 | Clinical Safety, Calibration, and Governance

High discrimination is not equivalent to calibrated clinical risk. A model can have a high ROC-AUC while producing probabilities that are systematically too high or too low, and its positive predictive value will change when disease prevalence changes. Because the archived test sets are balanced, the reported precision values are conditional on that artificial class distribution. They should not be presented as expected real-world post-test probabilities. Prospective deployment would require calibration assessment, prevalence-aware threshold selection, decision-curve analysis, and external validation in the intended population.

The same caution applies to the chatbot. A health assistant should not independently modify medications, interpret emergencies, or issue definitive diagnoses. WHO guidance on AI for health emphasizes protecting autonomy, promoting safety and public interest, ensuring transparency, fostering responsibility, and supporting inclusiveness and sustainability [8]. The system's safest role is therefore educational and navigational unless a future version is validated and governed under an appropriate clinical framework.

6.4 | Methodological Quality Against Modern Reporting Guidance

Table 4. Publication safeguards and remaining requirements for clinical translation.

Issue	What this paper does	Why it matters	Requirement before clinical use
Task definition	Reframes outputs as risk screening; removes unsupported GDM diagnostic claim.	Prevents mismatch between dataset outcome and clinical label.	Use a pregnancy-specific cohort with accepted GDM reference standard.
Resampling	Reports archived balanced outputs and flags unrecoverable resampling order.	Synthetic samples in a test set can inflate apparent generalization.	Split first; fit preprocessing/SMOTE only inside training folds.

Validation	Reports exact internal confusion matrices and ROC-AUC.	Internal performance does not establish transportability.	Independent external and prospective validation.
Calibration	Does not equate AUC with probability accuracy.	Clinical decisions require meaningful absolute risk.	Calibration-in-the-large, slope, calibration plots, and threshold analysis.
Fairness	Does not infer subgroup fairness from aggregate metrics.	Performance can differ by age, sex, geography, or socioeconomic status.	Pre-specified subgroup evaluation with adequate sample sizes.
Chatbot governance	Frames assistant as educational/non-diagnostic.	Generative systems can produce incorrect medical content.	Clinical content governance, logging, escalation, and safety evaluation.

Table 4 converts potential reviewer objections into explicit methodological boundaries. This approach is consistent with TRIPOD+AI and PROBAST+AI: the paper reports what the evidence supports, separates development quality from clinical applicability, and identifies the validation steps that would change the system from a proof-of-concept digital-health platform into a clinically deployable prediction tool [6,7].

7 | Limitations and Threats to Validity

The principal limitation is data provenance. Two predictive tasks rely on community-distributed public benchmarks, and the manuscript does not contain a complete patient-recruitment protocol or source-cohort description for those data. Their performance therefore cannot be assumed to represent Egyptian, Arab, pregnant, primary-care, or hospital populations. The BRFSS-derived data have better survey provenance, but they remain U.S. self-reported population data and therefore differ from a clinical cohort.

A second limitation concerns the archived evaluation pipeline. The first two test outputs are balanced, and the exact sequencing of train/test splitting, oversampling, and model fitting is not preserved in the manuscript. If resampling occurred before splitting, performance could be optimistic. This paper avoids concealing that possibility. The correct next evaluation is a leakage-safe pipeline with all transformations fit exclusively on training folds and a locked natural-prevalence test set.

Third, the study reports discrimination and class metrics but no calibration analysis, decision-curve analysis, external validation, or prospective clinical study. No claim is therefore made that the application can diagnose diabetes, exclude diabetes after a negative result, or select treatment. Fourth, the user-interface screenshots demonstrate implementation but not clinical efficacy, usability at scale, accessibility compliance, or patient-outcome improvement. Finally, the conversational assistant has not been evaluated here for factuality, harmful advice, or emergency escalation, so it remains an educational component rather than a validated clinical agent.

8 | Conclusion

This study presents an integrated mobile platform that combines diabetes risk screening, daily self-management support, Arabic-language educational assistance, and location-aware care navigation. The archived internal evaluations show strong discrimination for the three implemented tabular classifiers: the Random Forest general-risk model achieved 94.10% accuracy and ROC-AUC 0.9916; the SVC women-focused model achieved 90.29% accuracy and ROC-AUC 0.95; and the XGBoost health-indicator model achieved 94.99% accuracy and ROC-AUC 0.9900. These values are useful engineering evidence, but they are not presented as clinical validation.

The most important contribution of the publication version is the alignment of claims with evidence. The second dataset is no longer described as proof of gestational-diabetes diagnosis, the BRFSS model is treated as a health-indicator risk classifier rather than a physiological “daily score,” and all predictive outputs are

framed as screening signals that require professional confirmation. This disciplined scope makes the system more scientifically credible and creates a clear path for future work: leakage-safe re-evaluation, calibration, independent external validation, prospective testing in the target population, subgroup fairness analysis, and governed clinical evaluation of the conversational assistant. The architecture is already modular enough to accommodate those stronger evidence layers without redesigning the entire application.

Author Contribution

All authors contributed equally to this study.

Funding

This research received no external funding.

Data Availability

The study uses publicly available datasets cited in the manuscript: the Diabetes Prediction Dataset distributed on Kaggle [16], the 1,000-record Diabetes Prediction benchmark [17], and the Diabetes Health Indicators Dataset derived from BRFSS 2015 [18]. The numerical results reported here are reconstructed from the project's archived confusion matrices and ROC plots. No new patient-level dataset is released with this manuscript.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

- [1] I. Genitsaridi, P. Salpea, A. Salim, et al., "11th edition of the IDF Diabetes Atlas: global, regional, and national diabetes prevalence estimates for 2024 and projections for 2050," *The Lancet Diabetes & Endocrinology*, vol. 14, no. 2, pp. 149–156, 2026. doi:10.1016/S2213-8587(25)00299-2.
- [2] International Diabetes Federation, *IDF Diabetes Atlas*, 11th ed. Brussels, Belgium: International Diabetes Federation, 2025. <https://diabetesatlas.org>.
- [3] American Diabetes Association Professional Practice Committee for Diabetes, "2. Diagnosis and Classification of Diabetes: Standards of Care in Diabetes—2026," *Diabetes Care*, vol. 49, Suppl. 1, pp. S27–S49, 2026. doi:10.2337/dc26-S002.
- [4] R. Tarricone, F. Petracca, L. Svae, M. Cucciniello, and O. Ciani, "Which behaviour change techniques work best for diabetes self-management mobile apps? Results from a systematic review and meta-analysis of randomised controlled trials," *eBioMedicine*, vol. 103, 105091, 2024. doi:10.1016/j.ebiom.2024.105091.
- [5] M. Versluis et al., "Diabetes self-management education and support delivered by mobile health (mHealth) interventions for adults with type 2 diabetes—A systematic review and meta-analysis," *Diabetic Medicine*, vol. 42, no. 5, e70002, 2025. doi:10.1111/dme.70002.
- [6] G. S. Collins, K. G. M. Moons, P. Dhiman, et al., "TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods," *BMJ*, vol. 385, e078378, 2024. doi:10.1136/bmj-2023-078378.
- [7] K. G. M. Moons, J. A. A. Damen, T. Kaul, et al., "PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods," *BMJ*, vol. 388, e082505, 2025. doi:10.1136/bmj-2024-082505.
- [8] World Health Organization, *Ethics and Governance of Artificial Intelligence for Health: WHO Guidance*. Geneva: World Health Organization, 2021. ISBN 978-92-4-002920-0.
- [9] D. Ehrmann, N. Hermanns, K. Finke-Gröne, et al., "Efficacy of a Digital Diabetes Logbook for People With Type 1, Type 2, and Gestational Diabetes: Results From a Multicenter, Open-Label, Parallel-Group, Randomized Controlled Trial," *Journal of Diabetes Science and Technology*, vol. 19, no. 5, pp. 1362–1373, 2025. doi:10.1177/19322968241239870.
- [10] V. Mohan, S. Kalra, A. Augustine, and J. Kober, "Evaluating the Impact of mySugr Mobile Health Application on Glycemic Control in People with Diabetes Mellitus in India: A Real-World Data Analysis," *Diabetes Therapy*, vol. 16, no. 8, pp. 1681–1693, 2025. doi:10.1007/s13300-025-01768-x.
- [11] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, pp. 5–32, 2001. doi:10.1023/A:1010933404324.
- [12] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995. doi:10.1007/BF00994018.

- [13] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794. doi:10.1145/2939672.2939785.
- [14] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002. doi:10.1613/jair.953.
- [15] D. Luxen and C. Vetter, "Real-time routing with OpenStreetMap data," in Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, 2011, pp. 513–516. doi:10.1145/2093973.2094062.
- [16] Mustafa T., "Diabetes Prediction Dataset," Kaggle, public dataset, accessed 2026. <https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset>.
- [17] MrSimple07, "Diabetes Prediction," Kaggle, public dataset, accessed 2026. <https://www.kaggle.com/datasets/mrsimple07/diabetes-prediction>.
- [18] A. Teboul, "Diabetes Health Indicators Dataset," Kaggle; cleaned derivatives of the CDC Behavioral Risk Factor Surveillance System 2015, accessed 2026. <https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset>.

Disclaimer/Publisher's Note: The perspectives, opinions, and data shared in all publications are the sole responsibility of the individual authors and contributors, and do not necessarily reflect the views of Sciences Force or the editorial team. Sciences Force and the editorial team disclaim any liability for potential harm to individuals or property resulting from the ideas, methods, instructions, or products referenced in the content.