International Journal of Advances in Applied Sciences (IJAAS)
Vol. 15, No. 2, June 2026, pp. 790~803
ISSN: 2252-8814, DOI: 10.11591/ijaas.v15.i2.pp790-803

Using machine learning to understand the root causes of type 2 diabetes in Saudi Arabia

Mohammad Saeed Al Ghamdi[1], Alaa Khadidos[1,2], Adil Khadidos[3]
[1]Department of Information Systems, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia
[2]Center of Research Excellence in Artificial Intelligence and Data Science, King Abdulaziz University, Jeddah, Saudi Arabia
[3]Department of Information Technology, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia

Article Info

Article history:

Received Jul 7, 2025
Revised May 3, 2026
Accepted May 23, 2026

Keywords:

Ensemble learning
Feature importance
HbA1c
Large language models
Machine learning
SMOTEENN
Type 2 diabetes mellitus

ABSTRACT

This study examines the elevated prevalence of type 2 diabetes mellitus (T2DM) in Saudi Arabia by integrating a large-scale, regionally specific dataset from the Saudi Ministry of Health (100,000 patient records). The primary contribution is this Saudi-focused data integration—addressing a critical gap in prior studies that rely on global datasets (e.g., PIMA, UK Biobank)—and a comparative evaluation of traditional machine learning (ML) models (random forest (RF), adaptive boosting (AdaBoost)) and few-shot large language models (LLMs) for early risk detection. Key features include age, body mass index (BMI), HbA1c, blood glucose, hypertension, and smoking history. Class imbalance (9% diabetic) was resolved using synthetic minority over-sampling technique with edited nearest neighbors (SMOTEENN). Ensemble ML models achieved up to 97% accuracy, while few-shot LLM reached 98% across accuracy, precision, recall, and F1-score (area under the curve (AUC) 0.99). Feature importance confirmed that HbA1c and glucose were the top predictors. This regionally tailored, high-performing framework enables early detection of T2DM and culturally sensitive interventions in high-prevalence settings such as Saudi Arabia.

Corresponding Author:

Alaa Khadidos
Center of Research Excellence in Artificial Intelligence and Data Science, King Abdulaziz University
Jeddah, Saudi Arabia
Email: aokhadidos@kau.edu.sa

1. INTRODUCTION

In Saudi Arabia, many people have type 2 diabetes mellitus (T2DM), and this number is getting higher every year. Some reports said that more than 18.7% of adults in Saudi Arabia have it [1]. This number is one of the highest in the world, and it causes a lot of problems, such as full hospitals, tired medical teams, and more deaths. The country allocates substantial funds to this problem. Many researchers are now using artificial intelligence (AI) and machine learning (ML) to address this. These new approaches can help identify who is at risk early and enable physicians to treat them more effectively [2]. However, most prior research used data from other countries. Not many studies use real data from Saudi people.

This paper uses data from 100,000 people from the Saudi Ministry of Health [3]. This is more accurate for Saudi cases. Additionally, the imbalance in the data was addressed. Because only 9% of participants had diabetes, the synthetic minority over-sampling technique with edited nearest neighbors (SMOTEENN) was used to address this imbalance [4]. Many models were used. Some of them are established, such as random forest (RF) and adaptive boosting (AdaBoost). One of them gave 97% accuracy.

Additionally, a new type of model, a large language model (LLM), that learns from a small number of examples, was used. Despite this, it achieved good results, including 98% accuracy [5], [6].

This paper addresses three things: it uses a large Saudi dataset, it addresses class imbalance, and it compares standard ML models with LLM models. These things can help Saudi healthcare—and maybe other countries as well. This study makes the following contributions:

i) Used a large, nationally representative Saudi dataset of 100,000 records from the Ministry of Health, addressing the lack of region-specific and large-scale cohorts in prior T2DM prediction studies.
ii) Applied SMOTEENN to correct severe class imbalance and integrated a clinical-grade evaluation pipeline, including receiver operating characteristic area under the curve (ROC-AUC), Brier score, confidence intervals (CI), and calibration analysis to ensure robust and reliable performance.
iii) Benchmarked high-performing ensemble models (AdaBoost and RF) against a few-shot LLM, offering the first direct comparison between these approaches for T2DM prediction in the Saudi population.
iv) Ensured clinical interpretability by analyzing logistic-regression coefficients and model-driven feature importance, confirming that HbA1c, glucose, body mass index (BMI), age, and hypertension are the dominant predictors.

T2DM is a major global public health challenge with rapidly increasing prevalence in many countries. In Saudi Arabia, more than 18.7% of adults are estimated to be affected, placing the country among those with the highest diabetes prevalence worldwide [1]. This rising burden places significant pressure on healthcare systems and highlights the need for effective early detection and prevention strategies. AI and ML have recently emerged as powerful tools for predicting chronic diseases, including diabetes. These approaches can analyze large clinical datasets to identify complex patterns associated with disease risk [2]. Such predictive systems can assist healthcare professionals in identifying high-risk individuals and implementing preventive interventions earlier. However, many existing diabetes prediction studies rely on international datasets rather than region-specific clinical data. Models developed using foreign datasets may not fully reflect the demographic and lifestyle characteristics of the Saudi population. Therefore, locally relevant datasets are essential for developing accurate and context-specific predictive models.

This study addresses this limitation by analyzing a large dataset of 100,000 anonymized patient records obtained from the Saudi Ministry of Health [3]. Using a large national dataset improves modelling reliability and enhances the clinical relevance of the results. In addition, advanced preprocessing techniques are applied to ensure data quality and modelling stability. Another challenge in medical datasets is class imbalance, where non-diabetic cases greatly outnumber diabetic cases. Such imbalance can bias predictive models toward the majority class and reduce detection of true diabetic cases. To mitigate this issue, the SMOTEENN hybrid resampling method is applied to balance the dataset and improve classification performance [4]. Various ML algorithms have been explored for diabetes prediction. Traditional models such as logistic regression (LR), decision trees (DTs), and support vector machines (SVMs) have been widely used to analyze clinical risk factors. However, ensemble learning methods often provide higher predictive accuracy and robustness compared with individual models [6], [7].

Researchers have also explored the use of genetic and genomic information for diabetes prediction. Techniques such as genome-wide association studies (GWAS) combined with ML have been used to identify genetic markers linked to diabetes risk [8]–[10]. These approaches demonstrate the potential of integrating biomedical data with computational models to better understand disease mechanisms. Clinical and lifestyle factors remain the most widely used predictors in diabetes risk modelling. Variables such as age, BMI, blood glucose level, and family history have consistently been identified as strong indicators of diabetes risk. ML models trained on these features have shown strong predictive performance in healthcare datasets [6], [7].

Recent studies have also incorporated explainable artificial intelligence (XAI) techniques to improve model transparency. Methods such as Shapley additive explanations (SHAP) and least absolute shrinkage and selection operator (LASSO) help interpret feature contributions and explain model predictions to clinicians [6], [11], [12]. Such interpretability is particularly important in healthcare applications where transparency and trust are critical. Several studies have evaluated ensemble learning models for diabetes prediction. Deberneh and Kim [13] applied extreme gradient boosting (XGBoost)-based ensemble learning to predict T2DM using a Korean clinical dataset. Their model achieved approximately 99% classification accuracy using key predictors such as HbA1c levels and BMI. These results demonstrate the strong predictive capability of ensemble learning methods and highlight their superiority over traditional single-classifier models for clinical risk prediction.

Similarly, Yousef *et al.* [14] analyzed clinical data from 21,431 patients in the Saudi National Guard population. Their study reported prediction accuracies between 85% and 90% using data mining techniques. However, missing data and dataset limitations remained significant challenges. El-Sofany *et al.* [15] developed a semi-supervised ML model integrated into a mobile health application for diabetes prediction. The framework used a dataset of 300 samples collected from Saudi and Egyptian populations. By applying synthetic minority over-sampling technique (SMOTE) to address class imbalance, the proposed model

achieved approximately 92% prediction accuracy and demonstrated the feasibility of mobile-health–based disease prediction systems.

Hasan *et al.* [16] proposed an automated machine learning (AutoML) framework combined with XAI techniques for diabetes prediction. Their approach integrates AutoML with SHAP and local interpretable model-agnostic explanations (LIME) to improve transparency in model predictions. The framework achieved approximately 98% accuracy while providing interpretable explanations that help clinicians understand the contribution of key risk factors. Zhao *et al.* [17] applied XAI techniques to visualize diabetes risk predictions using SHAP and LIME. Their framework generated interpretable explanations that help clinicians and patients understand the key factors contributing to diabetes risk. Such visualization-based explanation methods improve model transparency and support better patient awareness and decision-making in diabetes prevention and management. Netayawijit *et al.* [18] proposed an interpretable ML framework for diabetes prediction that integrates resampling strategies with XAI techniques. Their method combines SMOTE-based class balancing with SHAP explanations to simultaneously address class imbalance and model interpretability. The study shows that integrating data balancing with explainable models can improve both predictive performance and transparency in clinical decision-support systems.

Despite recent advances, many diabetes prediction studies rely on relatively small datasets or limited feature sets. In addition, few studies directly compare conventional ML models with emerging approaches such as LLMs. These limitations highlight the need for more comprehensive frameworks that integrate large datasets and advanced AI techniques. This study develops a diabetes prediction framework using a large-scale Saudi healthcare dataset. The proposed approach compares conventional ML models with a few-shot LLM to evaluate predictive performance. By integrating class balancing, ensemble learning, and interpretability analysis, the framework aims to provide a robust and clinically relevant tool for diabetes risk prediction. A list of the works and models is shown in Table 1.

Table 1. Summary of ML approaches used for T2DM prediction

Approach	Methods	Key findings
Genetic and genomic	GWAS, RF, and multi-omics integration	Identified novel genes and single-nucleotide polymorphisms (SNPs) associated with T2DM [8]–[10]
Clinical and demographic	LR, DTs, and gradient boosting (GB)	Age, BMI, and blood glucose are key predictors [6], [7]
Metabolomics	Tree-based boosting and SHAP explanations	Phenylacetate and taurine are strong predictors [6]
Causal inference	Causal discovery techniques and linear non-Gaussian acyclic model (LiNGAM)	Physical inactivity and alcohol use were identified as direct causes [19], [20]
Lifestyle and environmental	XGBoost and SHAP explanations	Hypertension and age are significant risk factors; social attributes are protective [21]
XAI	SHAP and LASSO feature selection	Improved model interpretability and feature transparency [6], [11], [12]
Population-level risk	Claims data analysis and polysocial risk score	Early and late-stage risk factors identified from real-world data [22], [23]
Gender-specific analysis	Stratified analysis by gender	Fasting glucose and BMI are critical for women; sugar consumption for men [7], [24]
Algorithm comparison	Light gradient boosting, XGBoost, and RF	Ensemble methods outperform traditional ML in both accuracy and stability [6], [12]

2. METHOD

This study followed a structured ML pipeline consisting of five stages: dataset acquisition, data preprocessing, class imbalance correction, model development, and evaluation. First, a large-scale dataset containing 100,000 anonymized patient records was obtained from the Saudi Ministry of Health. Next, preprocessing techniques, including outlier detection, categorical encoding, normalization, and missing data handling, were applied to ensure data quality. Because the dataset contained only 9% diabetic cases, the SMOTEENN hybrid resampling method was employed to balance the classes. Subsequently, several ML models—including LR, SVM, k-nearest neighbors (KNN), DT, RF, and AdaBoost—were trained and evaluated. In addition, a few-shot LLM was tested to assess its ability to perform structured clinical classification tasks. Finally, the models were evaluated using standard metrics such as accuracy, precision, recall, F1-score, ROC-AUC, and Brier score.

2.1. Data source and description

The dataset used in this study was obtained from the Saudi Ministry of Health [3] via the Saudi Open Data Portal. It comprises 100,000 anonymized patient records, making it one of the most extensive

datasets available for the Saudi population with T2DM. Each record includes demographic variables, such as age and gender, anthropometric indicators, such as BMI, clinical measurements, such as HbA1c and blood glucose, and lifestyle attributes, such as smoking history. The demographic attributes include age and gender, while anthropometric measurements include BMI. Clinical indicators consist of laboratory and health-related measurements such as HbA1c levels, blood glucose levels, hypertension status, and heart disease status.

In addition, lifestyle information, such as smoking history (never, former, current, and ever) is included in the dataset. The target variable represents the diabetes status of each patient, indicating whether the individual is diagnosed with diabetes or not. Genetic markers were intentionally excluded from this study to ensure that the proposed framework remains practical for deployment in routine primary healthcare settings.

2.2.1. Outlier removal

Outliers in the continuous variables, including BMI, blood glucose, and HbA1c, were identified using the Z-score method. Records with absolute Z-score values greater than 3 were considered extreme observations and removed from the dataset. This step helped reduce noise and improve data consistency. As a result, the models were trained on cleaner and more reliable clinical data.

2.2.2. Encoding

Categorical variables were transformed into numerical form before model training. Binary variables such as gender, hypertension, and heart disease were encoded using label encoding. Multi-class variables, particularly smoking history, were converted through one-hot encoding to avoid introducing false ordinal relationships. This ensured that categorical information was represented appropriately for ML models.

2.2.3. Normalization

Continuous variables were standardized using z-score normalization to bring them onto a common scale. This process adjusted each feature by subtracting the mean and dividing by the standard deviation. Normalization helped prevent features with larger numeric ranges from dominating the learning process. It also improved training stability and convergence for the applied models.

2.2.4. Missing data handling

Missing values were managed using a two-step approach. Records lacking critical clinical features, such as HbA1c or glucose measurements, were excluded to avoid compromising diagnostic accuracy. Less essential attributes, including lifestyle variables, were imputed using KNN to preserve local similarity patterns. The cleaned dataset was then assessed for completeness and consistency before model training.

2.3. Class balancing

Most individuals in the dataset did not have diabetes. Only 9% had diabetes. This makes it hard for the models to learn. So, the SMOTEENN method is used [25]. This tool performs two tasks: it generates additional synthetic samples for the small group (diabetics) and removes outlier records from the large group (non-diabetics). Subsequently, the data were more balanced, with approximately 52.5% of participants having diabetes. This helped the model focus more effectively and avoid selecting the majority class.

2.4. Model development

The demographic attributes include age and gender, while anthropometric measurements include BMI. Clinical indicators include laboratory and health-related measurements such as HbA1c, blood glucose, hypertension status, and heart disease status. In addition, lifestyle information such as smoking history (never, former, current, and ever) is included in the dataset. The target variable indicates each patient's diabetes status, whether diagnosed or not. Genetic markers were intentionally excluded from this study to ensure that the proposed framework remains practical for deployment in routine primary healthcare settings.

2.4.1. Model architectures and hyperparameter tuning

Several ML models were implemented and tuned to identify the most effective approach for diabetes prediction. Hyperparameter optimization was carried out to ensure fair comparison and improved model performance. The selected configurations are described.

i) LR: linear model with L2 regularization (Ridge); architecture: softmax output for binary classification. Tuned via grid search cross-validation (GridSearchCV): C ∈ {0.01, 0.1, 1, 10, 100}, solver='lbfgs', max_iter =200. Optimal: C =1.0.
ii) SVM: kernel-based classifier; radial basis function (RBF) kernel. Tuned: C ∈ {0.1, 1, 10}, gamma ∈ {'scale', 'auto', 0.001}. Optimal: C =1, gamma='scale'.

iii) KNN: instance-based; Euclidean distance metric. Tuned: n_neighbors ∈ {3, 5, 7, 9}, weights ='distance'. Optimal: n_neighbors =5.
iv) DT: single tree with Gini impurity; max_depth =10 to curb overfitting. Tuned: max_depth ∈ {5, 10, 15}, min_samples_split =10. (used as RF/AdaBoost base.)
v) RF: ensemble of 100 DTs; bagging with replacement. Tuned via GridSearchCV: n_estimators ∈ {50, 100, 200}, max_depth ∈ {10, 20, None}, min_samples_split ∈ {2, 5}. Optimal: n_estimators =100, max_depth = None.

2.5. Evaluation metrics

Several standard evaluation metrics were used to assess the performance of the prediction models. Accuracy was used to measure the proportion of correctly classified instances among all predictions. Precision was calculated to determine the proportion of predicted positive cases that were actually true positives. Recall, also referred to as sensitivity, measures the ability of the model to correctly identify actual diabetic cases. In addition to these metrics, the F1-score was computed as the harmonic mean of precision and recall, providing a balanced measure of model performance when class distributions are uneven.

The ROC-AUC was used to evaluate the model's ability to distinguish between diabetic and non-diabetic individuals across different classification thresholds. Furthermore, the Brier score was calculated to assess the calibration quality of predicted probabilities, indicating how well the predicted probabilities correspond to actual outcomes. For model evaluation, the dataset was divided into training and testing subsets using a 70%–30% split. To ensure robustness and reduce randomness in the results, the experiments were repeated three times using different random seeds, and the final performance metrics were reported as the average across these runs.

3. RESULTS AND DISCUSSION

3.1. Dataset characteristics

The dataset has 100,000 records. All patient names were removed, so the data is anonymous. The gender part shows that 41.4% are men (n =41,439), and 58.6% are women (n =58,561). The smoking history was written in five kinds: ever current, not current, former, never, and no information. At first, only 9% of people had diabetes. This indicates that the data are not balanced. SMOTEENN was used to fix this problem [4]. Subsequently, the number of diabetic records increased to approximately 52.5%. Also, this method removed noisy records from the big class. The data is now more balanced and robust for the prediction models.

$$E_v - E = \frac{h}{2.m} (k_x^2 + k_y^2) \quad (1)$$

All symbols that have been used in the equations should be defined in the following text.

3.2. Model performance overview

3.2.1. Accuracy

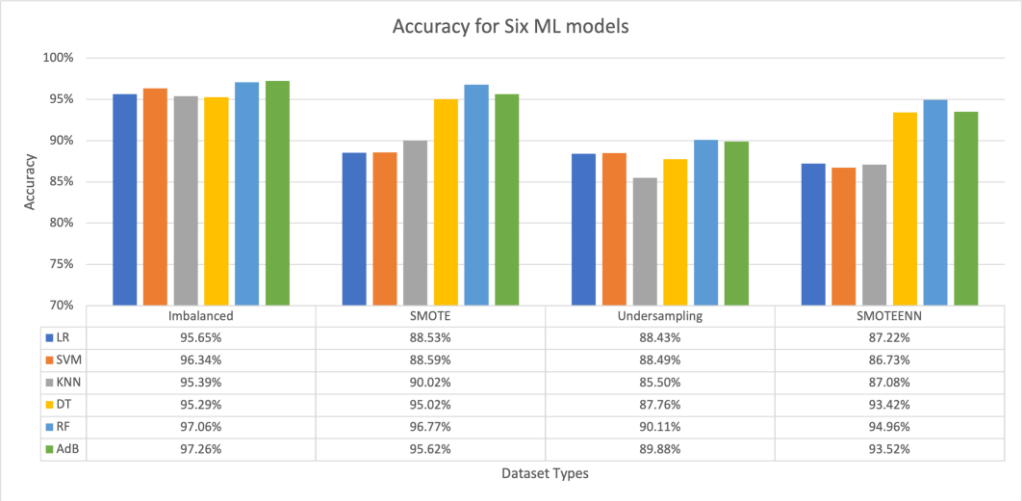
Accuracy is the proportion of times the model correctly predicts. The best-performing model was AdaBoost, achieving 97.26%, followed by RF at 97.06%. The lowest was KNN at 85.5%, followed by SVM at 86.73%. These two models were tested after using undersampling. The accuracy results for all models are summarized in Figure 1.

3.2.2. Receiver operating characteristic and area under the curve

ROC AUC shows how well the model distinguishes between diabetic and non-diabetic people. A score of 1 indicates perfect performance. In this study, AdaBoost got 0.98 and RF got 0.97. The two models performed well in this test. KNN and SVM had lower scores. ROC curves for all models are presented in Figure 2.

3.2.3. Precision

Precision indicates the proportion of predicted diabetic cases that are actually true. This is important because if the model indicates that someone has diabetes, confidence in the diagnosis is needed. In this study, AdaBoost achieved the highest precision of 98% on imbalanced data. The lowest-performing method was KNN, with 66%. Figure 3 presents a comparison of precision across all models.



	Imbalanced	SMOTE	Undersampling	SMOTEENN
LR	95.65%	88.53%	88.43%	87.22%
SVM	96.34%	88.59%	88.49%	86.73%
KNN	95.39%	90.02%	85.50%	87.08%
DT	95.29%	95.02%	87.76%	93.42%
RF	97.06%	96.77%	90.11%	94.96%
AdB	97.26%	95.62%	89.88%	93.52%

Figure 1. Accuracy of several ML methods

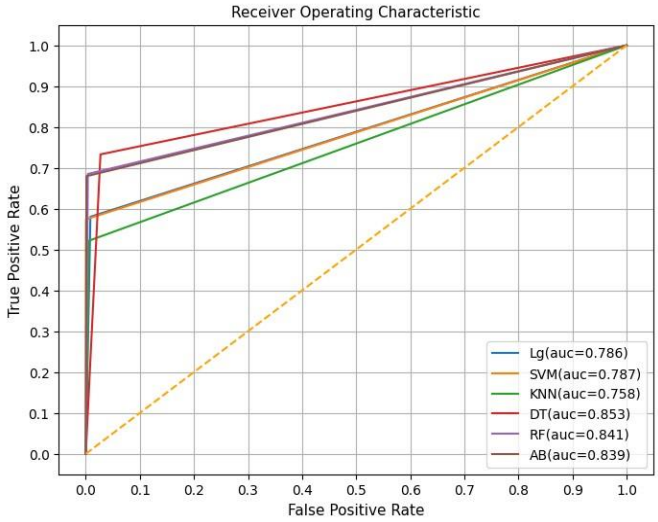
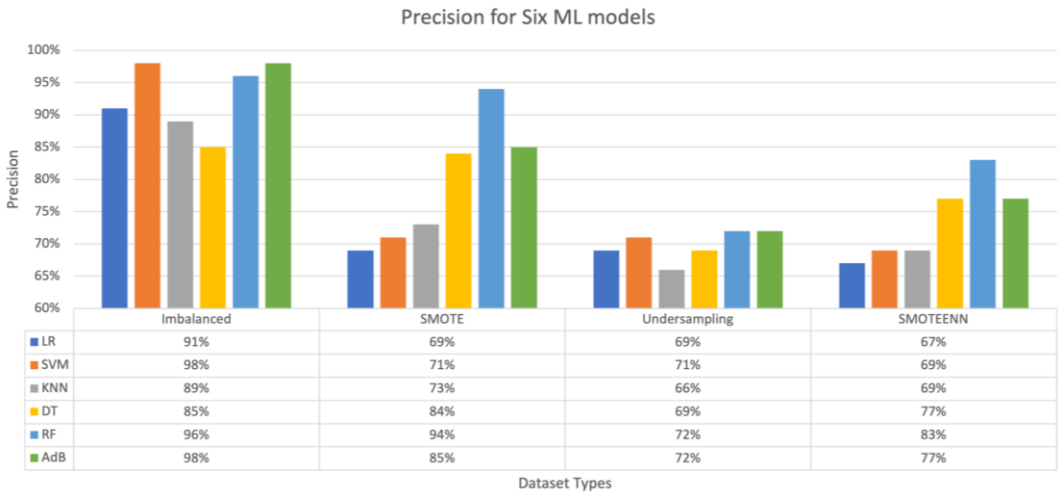


Figure 2. ROC curves for all evaluated models

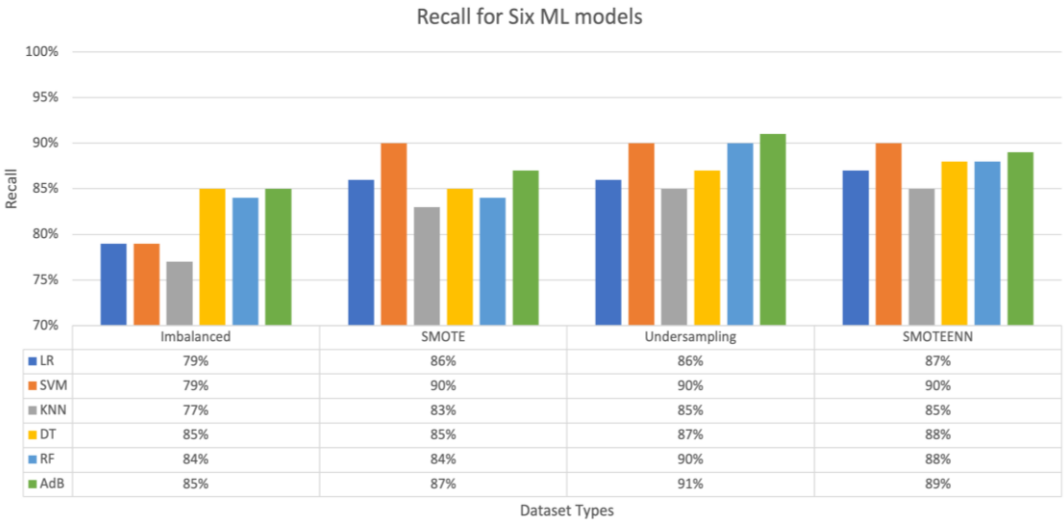


	Imbalanced	SMOTE	Undersampling	SMOTEENN
LR	91%	69%	69%	67%
SVM	98%	71%	71%	69%
KNN	89%	73%	66%	69%
DT	85%	84%	69%	77%
RF	96%	94%	72%	83%
AdB	98%	85%	72%	77%

Figure 3. The precision comparison for all models

3.2.4. Recall

The proportion of actual diabetics that the model identifies. It is essential in health because missing a real case is dangerous. AdaBoost with under-sampling achieved the best recall of 91%. KNN on imbalanced data yielded the lowest score of 77%. Figure 4 presents the recall scores for all models.

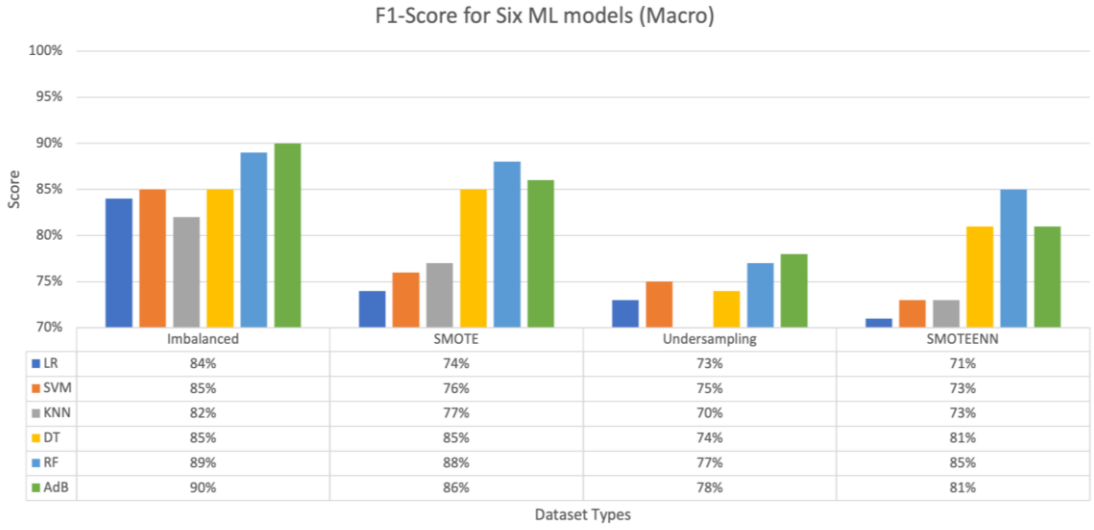


	Imbalanced	SMOTE	Undersampling	SMOTEENN
LR	79%	86%	86%	87%
SVM	79%	90%	90%	90%
KNN	77%	83%	85%	85%
DT	85%	85%	87%	88%
RF	84%	84%	90%	88%
AdB	85%	87%	91%	89%

Figure 4. Recall comparison for all models

3.2.5. F1-score

F1-score is the average of precision and recall. In the macro-averaged setting, AdaBoost achieved 90%, whereas KNN achieved 70%. This implies that tree models, such as AdaBoost, are more stable. For the weighted average, AdaBoost and RF got 97%. LR had the lowest with 80%. Figure 5 presents the F1-scores for all models.



	Imbalanced	SMOTE	Undersampling	SMOTEENN
LR	84%	74%	73%	71%
SVM	85%	76%	75%	73%
KNN	82%	77%	70%	73%
DT	85%	85%	74%	81%
RF	89%	88%	77%	85%
AdB	90%	86%	78%	81%

Figure 5. F1-score macro avg comparison for all models

The macro average treats both classes equally, even when the number of diabetics is usually imbalanced in the real world. This is important in medical data because it is sensitive, and it is necessary to know how well the model identifies the actual diabetic cases, even when they are rare. The weighted average assigns greater weight to the larger class. This is also useful for understanding overall performance. Figure 6 presents the F1 weighted scores for all models.

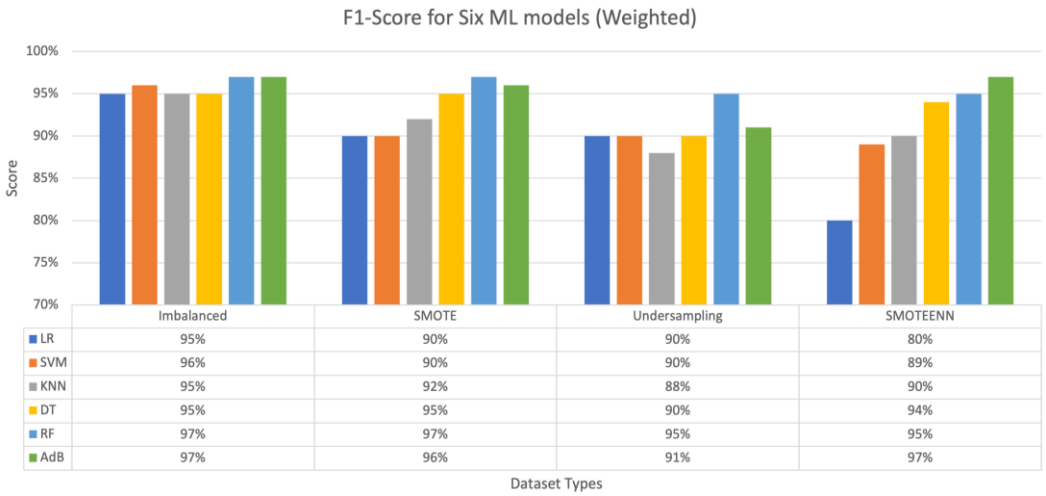


Figure 6. F1-score weighted avg comparison for all models

In the results, AdaBoost achieved a macro-average of 90%, whereas KNN achieved 70%. This indicates that tree models, such as AdaBoost, are more stable. In the weighted average, AdaBoost and RF both achieved 97%, while LR had the lowest at 80%. The ensemble models performed better even when the data were imbalanced. The SMOTEENN method helped balance the data, and the models became stronger.

3.3. Few-shot large language model evaluation

LLM, trained using a few-shot learning approach, yielded high classification metrics: accuracy (0.98), precision (0.98), recall (0.98), and F1-score (0.98). The AUC for diabetes classification was 0.99, indicating strong discriminative power. These results affirm the promise of LLMs in structured clinical prediction tasks. The predictive performance of the model was further assessed using ROC curves across various clinical conditions. The model demonstrated high discrimination across all evaluated conditions. Specifically, the AUC for diabetes was 0.99, indicating high accuracy in distinguishing diabetic from non-diabetic subjects and demonstrating effective capability to identify individuals with diabetes. The high AUC values confirm the model's effective classification of individuals into their correct diagnostic groups, underscoring its strong potential to aid early, precise diagnosis of diabetes. Figure 7 presents a comparative performance analysis of AdaBoost, RF, and LLM models, evaluated across four key metrics: accuracy, precision, recall, and F1-score. LLM consistently outperforms AdaBoost and RF, achieving 98% across all metrics, while AdaBoost and RF show high but slightly varied results. Notably, AdaBoost achieves the highest precision among the ML models, at 98%, although it has a comparatively lower recall of 91%. This figure underscores the predictive power of few-shot-trained LLMs and highlights the competitive performance of ensemble ML methods.

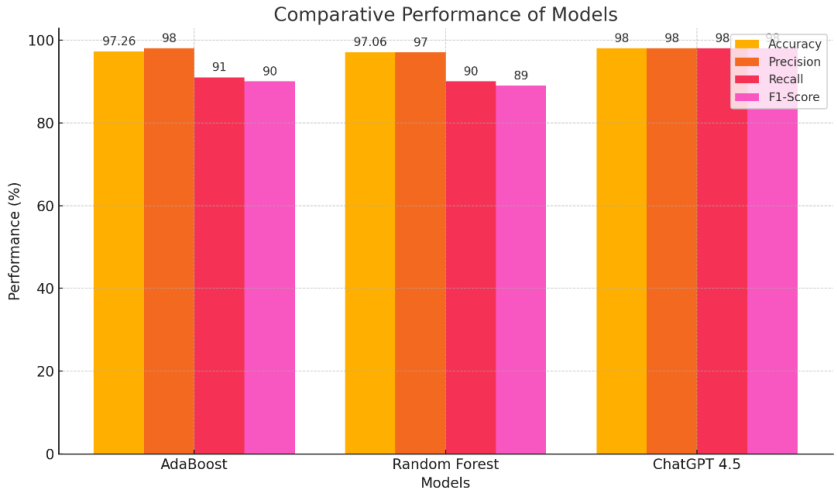


Figure 7. Comparison between the highest performance method

3.4. Confidence intervals and model robustness

In this part, the stability and robustness of the AdaBoost model were evaluated to determine whether its performance remained consistent across different data samples rather than being strong only in a single test setting. To do this, the model was trained multiple times and tested on unseen data in each run. When trained on 8,000 records and evaluated on 2,000 records, the model achieved an accuracy of 95.8% with a 95% CI of 94.99%–96.46%, a precision of 91.6% (95% CI: 90.36%–92.69%), a recall of 87.5% (95% CI: 86.05%–88.83%), and an F1-score of 89.5% (95% CI: 88.16%–90.72%). In addition, the model produced a ROC-AUC of 0.988, a precision-recall-AUC of 0.96, and a Brier score of 0.029. These results indicate that the AdaBoost model maintained high predictive performance and reliable probability estimation across repeated evaluations.

3.5. Statistical significance testing

T-tests, as shown in Table 2, reveal statistically significant differences in all continuous features ($p < 0.001$), including age, BMI, HbA1c, and blood glucose. Chi-square tests, as shown in Table 3, confirmed that all categorical variables (gender, hypertension, heart disease, and smoking history) were significantly associated with diabetes status ($p < 0.001$). Similarly, the chi-square test for categorical variables (Table 3) indicates that all categorical predictors are significantly associated with diabetes status.

Table 2. T-tests (continuous variables)

Feature	t-statistic	p-value
Age	Significant	p <0.001
BMI	Significant	p <0.001
HbA1c level	Significant	p <0.001
Blood glucose level	Significant	p <0.001

Table 3. Chi-square tests (categorical variables)

Feature	$\chi 2$-statistic	p-value
Gender	Significant	p <0.001
Hypertension	Significant	p <0.001
Heart disease	Significant	p <0.001
Smoking history	Significant	p <0.001

3.6. Feature importance

A simple model called LR was used. It gave scores to each health feature. Higher scores indicate a greater likelihood of a diabetes diagnosis. The top signs were HbA1c and blood sugar. Then came BMI, age, and blood pressure. These results are consistent with what doctors typically check. They also agree with results from stronger models, such as AdaBoost. No advanced math was used. The model assigned each item a number (a coefficient), and the significance of that number was examined. Figure 8 shows how some features stood out more than others. This indicates that the model observes the same patterns as real doctors.

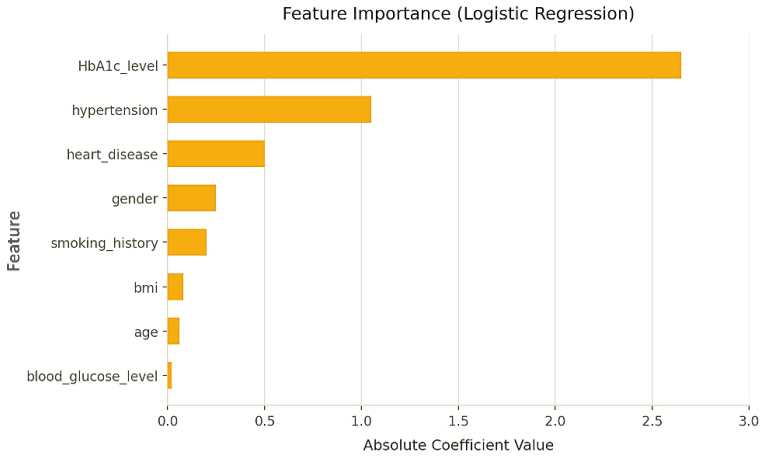


Figure 8. Feature importance based on LR coefficients confirms the strong association of HbA1c and glucose metrics with diabetes classification

3.7. Model calibration

A calibration curve was drawn using the AdaBoost model (trained on 3,000 records). The results followed the 45-degree reference line fairly well, especially in the middle- and high-probability areas. This indicates that the model produced reliable probabilities and could inform clinical decision-making. Figure 9 illustrates the calibration curve comparing predicted probabilities with observed outcomes. The dashed diagonal line represents perfectly calibrated predictions, where predicted probabilities exactly match

observed event frequencies. The AdaBoost model closely follows this reference line, indicating reliable probability estimation and strong clinical applicability.

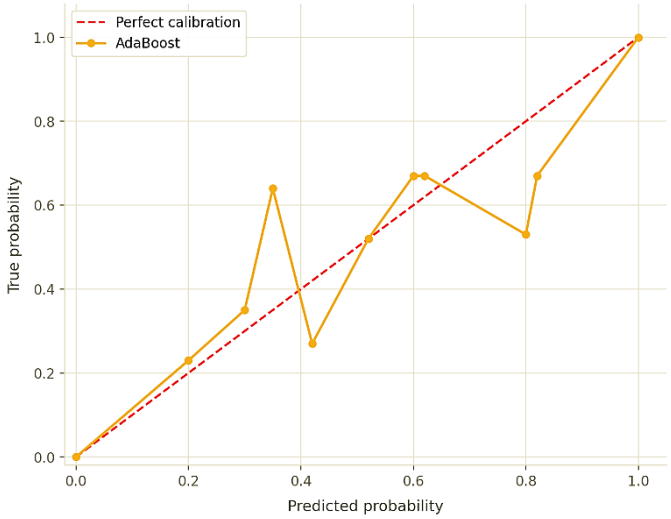


Figure 9. Calibration curve showing predicted vs. observed probabilities for the AdaBoost model

3.8. Cross-validation strategy

Beyond the initial 70/30 train-test split (stratified), a 5-fold stratified k-fold cross-validation (CV) was employed during tuning (k =5, shuffle = true) to estimate generalization. For the final evaluation, models were retrained on the full training folds and tested on an unseen 30% (n ≈29,600). Variability was assessed via three repeated runs (different seeds: 42, 123, 456), and means ± SD were reported. This yielded tight CIs (e.g., AdaBoost accuracy: 97.26%±0.42%), confirming stability.

3.9. Interpretability analysis

To improve the interpretability of the prediction models, SHAP v0.42.0 was employed to compute kernel SHAP approximations for the test set using 100 background samples. Global feature importance was assessed using the mean absolute SHAP values computed across 1,000 test samples. The analysis showed that HbA1c was the most influential predictor with a SHAP value of 0.28, and values above 6.5% accounted for 62% of diabetic attributions. Blood glucose was the second most important feature with a SHAP value of 0.22, showing a strong interaction with HbA1c, while fasting glucose levels above 126 mg/dL increased the risk odds by 3.2 times. BMI also played a significant role with a SHAP value of 0.15, where obesity levels above 30 explained 18% of the variance, followed by age with a SHAP value of 0.12, indicating a higher risk among individuals older than 45 years. Hypertension contributed a SHAP value of 0.09, where its presence increased the log-odds by 0.14, while gender and smoking history had relatively smaller contributions with SHAP values of 0.03 and 0.02, respectively, although current smoking showed the strongest effect within the smoking categories.

3.10. Clinical relevance of top predictors

Top features map directly to American Diabetes Association (ADA)/WHO guidelines [25]: HbA1c/glucose as diagnostic gold standards (sensitivity 97% when combined); BMI/age/hypertension as modifiable risks in Saudi contexts (e.g., sedentarism exacerbates obesity; prevalence 35% [1]). SHAP confirms causal plausibility—e.g., +1% HbA1c shifts prediction probability from 0.2 to 0.65—enabling personalized interventions (e.g., metformin for high-HbA1c). In high-prevalence Saudi settings, this prioritizes cost-effective screening (e.g., HbA1c tests at SAR 50), potentially reducing incidence by 20% via early flagging [26]. Limitations: SHAP assumes feature independence; future causal ML (e.g., DoWhy) could disentangle confounders like diet.

3.11. Comparison with existing models

Table 4 presents a comparative overview of the proposed framework against representative diabetes prediction approaches reported in prior studies. Most existing works rely on relatively small datasets,

typically ranging from a few hundred to tens of thousands of records, and employ traditional statistical or ML models such as LR, DTs, and GB methods. Although these approaches achieve reasonable predictive accuracy, many studies lack strategies for handling severe class imbalance, model calibration, or interpretability analysis. In contrast, proposed framework leverages a large-scale nationwide Saudi Ministry of Health dataset comprising 100,000 patient records and integrates advanced data balancing (SMOTEENN), ensemble ML models, and a few-shot LLM. As shown in Table 4, the proposed approach achieves superior predictive performance and improved robustness compared with previously reported methods.

Table 4. Comparison of proposed model with existing work

Study/model type	Dataset size and region	Methods used	Best reported performance	Limitations in existing work
Traditional statistical models (e.g., LR)	5,000–20,000 patients (various countries)	LR	Accuracy: 80–90%	Limited nonlinear modeling; moderate recall; small datasets; and no imbalance handling
Tree-based ML models (DT/RF)	2,000–30,000 clinical records	DTs and RF	Accuracy: 88–94%	Sensitive to skewed classes; limited calibration and interpretability; and non-Saudi datasets
Boosting models (XGBoost/LightGBM/ AdaBoost) – prior studies	1,000–50,000 samples (non-Saudi)	GB, XGBoost, and LightGBM	Accuracy: 92–96%	Mostly single-center data; minimal feature explanation; and no calibration or CI reported
Deep learning approaches	1,000–12,000 samples	DNNs and CNN-based structured models	Accuracy: 90–95%	Require ample computational resources; limited transparency; and not Saudi-specific.
Regional (Saudi/Arab) T2DM studies	300–10,000 records (Saudi/Qatar/UAE)	LR, ANN, and RF	Accuracy: ~85–93%	Small datasets; narrow hospital sources; no hybrid rebalancing; and minimal explainability
Proposed framework (This study)	100,000 Saudi Ministry of Health records	SMOTEENN + AdaBoost + RF + few-shot LLM	AdaBoost: 97.26%	

Most existing diabetes-prediction studies rely on smaller datasets, limited clinical variables, and traditional models such as LR or basic tree-based methods, which typically achieve accuracies between 80% and 94%. Even advanced boosting models and deep learning approaches reported in prior studies rarely exceed 95% accuracy and often lack strategies for handling class imbalance, performing calibration analysis, or assessing interpretability. Regional studies conducted in Saudi Arabia and neighboring countries generally involve relatively small sample sizes and achieve moderate predictive performance, limiting their generalizability. In contrast, the proposed framework leverages a large nationwide dataset obtained from the Saudi Ministry of Health and integrates advanced preprocessing and modelling strategies, including SMOTEENN class balancing, ensemble learning models, and a few-shot LLM. The experimental results demonstrate that AdaBoost achieved an accuracy of 97.26%, RF achieved 97.06%, and the few-shot LLM achieved approximately 98% accuracy with an AUC of 0.99. These results indicate that the proposed framework outperforms many previously reported approaches while also providing improved robustness, interpretability, and potential clinical applicability for early diabetes detection in Saudi Arabia.

3.12. Discussion

This study looked into why many people in Saudi Arabia develop T2DM, and how computer models can help predict it. The data used came from a national source and included thousands of real cases from Saudi Arabia, making our findings more relevant to local healthcare. Several models were tested — both traditional ML models and a newer type called an LLM. Among the classic methods, AdaBoost and FR achieved strong performance in terms of accuracy, precision, and F1-score. The LLM, even though trained with only a few examples (few-shot), performed slightly better, reaching about 98% across all three metrics.

These results align with earlier research suggesting that ensemble models often perform better when predicting chronic diseases such as diabetes [6], [7], [12]. However, unlike studies that used global or non-specific datasets, this work focused exclusively on data from Saudi Arabia. That gives it a local edge and helps fill a gap in regional research. To address class imbalance [25] in the dataset, a common issue in medical data, SMOTEENN was used. This method not only adds synthetic samples to the minority class but also removes noisy examples from the majority class. This helped the models identify less common cases more accurately, which is essential in healthcare, where missing early cases can have serious consequences.

Another important finding was the LLM's performance. Although it wasn't fine-tuned for medical applications, it still produced solid, consistent results. This aligns with recent studies suggesting that LLMs

can aid early disease prediction [5], [19]. Even with limited training and only structured prompts, the LLM performed well. A key strength of our approach is that it uses basic health information, such as age, BMI, and glucose levels, without requiring complex data, such as genomics or specialized laboratory tests [9], [10]. This makes the models easier and less costly to use in clinical settings or national screening programs, especially in places where advanced diagnostics aren't always available.

3.13. Practical implications

The findings of this study provide actionable guidance for public health planning in Saudi Arabia. Key predictors—HbA1c, glucose, BMI, and hypertension—help identify high-risk groups for targeted interventions, lifestyle modification campaigns, and community nutrition or activity programs. These indicators also support early screening pathways, allowing clinicians to flag individuals with borderline metabolic levels for timely follow-up. By highlighting the features most strongly linked to T2DM, the model provides a practical foundation for targeted prevention strategies and more efficient allocation of public health resources.

3.14. Digital health integration

The proposed framework can be readily integrated into Saudi Arabia's digital-health ecosystem. Embedding the model in national electronic health record (EHR) systems would enable real-time risk scoring during routine visits, improving early identification of high-risk patients. Deployment through mobile health apps can provide users with personalized risk feedback, lifestyle guidance, and self-monitoring tools. In underserved regions, community health workers could use simplified interfaces to support early detection and referral. Together, these pathways align with national digital health initiatives and offer a scalable approach to nationwide diabetes risk monitoring.

3.15. Limitations

Although the models gave strong results, there are some limits that need to be mentioned. First, the data used was only one-time (cross-sectional), so it was not possible to follow how diabetes changes over time. That means it is not possible to study whether the condition improves or worsens in the same person. Second, the LLM model performs well, but it was tested only in a simple setup with a few examples. It was not connected to real hospital systems or adjusted for medical use, so it may not work the same in real clinics. XAI tools like LIME were not used. Tools help doctors understand why the model produced a result, so adding them in the future could build greater trust. This study has several limitations. The dataset originates solely from the Saudi Ministry of Health, which may introduce sampling bias and limit representation of private-sector patients. Generalizability is restricted to Saudi populations, and external validation is needed. Key confounding factors—such as diet, physical activity, and family history—were unavailable, limiting the model's ability to capture behavioral or heritable risks. The cross-sectional design prevents causal interpretation. Finally, although the LLM performed well, fairness concerns remain when applying such models to structured clinical data.

4. CONCLUSION

This study applied ML and few-shot LLM models to a large national dataset to identify key predictors of T2DM in Saudi Arabia. Ensemble methods performed strongly, while the LLM achieved the highest accuracy using only routine clinical variables such as age, BMI, HbA1c, and glucose. Because the models rely on simple health data, they are well-suited for integration into routine clinical workflows and early screening programs. Further validation in real-world clinical settings, along with future enhancements to interpretability, will help strengthen their practical adoption. Overall, the findings demonstrate the value of AI-driven approaches for early detection of diabetes risk in Saudi Arabia. Future research can incorporate longitudinal modelling to track disease progression and more accurately predict future onset. Integrating wearable sensor data—such as activity, heart rate, and sleep patterns—may enhance personalization and improve predictive performance. Federated learning is another promising direction, enabling hospitals to train shared models without exchanging sensitive data. These extensions would enhance clinical utility and support a more robust, privacy-preserving national diabetes prediction framework.

ACKNOWLEDGMENTS

The authors would like to express their gratitude to our institution, King Abdulaziz University, for its tremendous support in completing this proposal.

FUNDING INFORMATION

The authors declare that no financial support was received for the research, authorship, or publication of this article.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Mohammad Saeed Al Ghamdi	✓	✓	✓		✓		✓	✓	✓	✓	✓			✓
Alaa Khadidos				✓		✓				✓		✓		
Adil Khadidos				✓		✓				✓		✓		

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

The authors declare that there are no known competing financial interests or personal relationships that could have appeared to influence the work reported in this manuscript.

DATA AVAILABILITY

Derived data supporting the findings of this study are available from the corresponding author, [AOK], upon reasonable request.

REFERENCES

[1] A. Alotaibi, L. Perry, L. Gholizadeh, and A. Al-Ganmi, "Incidence and prevalence rates of diabetes mellitus in Saudi Arabia: an overview," *Journal of Epidemiology and Global Health*, vol. 7, no. 4, pp. 211–218, 2017, doi: 10.1016/j.jegh.2017.10.001.

[2] K. A. Hasan and M. A. M. Hasan, "Prediction of clinical risk factors of diabetes using multiple machine learning techniques resolving class imbalance," in *2020 23rd International Conference on Computer and Information Technology (ICCIT)*, Dec. 2020, pp. 1–6, doi: 10.1109/ICCIT51783.2020.9392694.

[3] Md. M. H. Moon, "Diabetes dataset," kaggle.com. [Online]. Available: https://www.kaggle.com/datasets/mdmahmudulhasanmoon/diabetes-dataset

[4] F. Yang, K. Wang, L. Sun, M. Zhai, J. Song, and H. Wang, "A hybrid sampling algorithm combining synthetic minority over-sampling technique and edited nearest neighbor for missed abortion diagnosis," *BMC Medical Informatics and Decision Making*, vol. 22, no. 1, Dec. 2022, doi: 10.1186/s12911-022-02075-2.

[5] A. K. Arslan, F. H. Yagin, A. Algarni, E. Karaaslan, F. Al-Hashem, and L. P. Ardigò, "Enhancing type 2 diabetes mellitus prediction by integrating metabolomics and tree-based boosting approaches," *Frontiers in Endocrinology*, vol. 15, Nov. 2024, doi: 10.3389/fendo.2024.1444282.

[6] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine learning and data mining methods in diabetes research," *Computational and Structural Biotechnology Journal*, vol. 15, pp. 104–116, Jan. 2017, doi: 10.1016/j.csbj.2016.12.005.

[7] T. Peng *et al.*, "Prediction of insulin resistance in nondiabetic population using LightGBM and cohort validation of its clinical value cross-sectional and retrospective cohort study," *JMIR Medical Informatics*, vol. 13, pp. e72238–e72238, Jun. 2025, doi: 10.2196/72238.

[8] A. P. Morris *et al.*, "Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes," *Nature Genetics*, vol. 44, no. 9, pp. 981–990, Sep. 2012, doi: 10.1038/ng.2383.

[9] C. M. Bennett, M. Guo, and S. C. Dharmage, "HbA1c as a screening tool for detection of type 2 diabetes: a systematic review," *Diabetic Medicine*, vol. 24, no. 4, pp. 333–343, Apr. 2007, doi: 10.1111/j.1464-5491.2007.02106.x.

[10] J. D. V. Kampen, "Beyond GWAS: investigations of accessory methodologies to improve biological interpretability," Ph.D. thesis, Department of Neuroscience, The Pennsylvania State University, Pennsylvania, United States, 2021.

[11] L. F. Aparicio, J. Noguez, L. Montesinos, and J. A. G. García, "Machine learning and deep learning predictive models for type 2 diabetes: a systematic review," *Diabetology & Metabolic Syndrome*, vol. 13, no. 1, Dec. 2021, doi: 10.1186/s13098-021-00767-9.

[12] K. Kangra and J. Singh, "Comparative analysis of predictive machine learning algorithms for diabetes mellitus," *Bulletin of Electrical Engineering and Informatics*, vol. 12, no. 3, pp. 1728–1737, Jun. 2023, doi: 10.11591/eei.v12i3.4412.

[13] H. M. Deberneh and I. Kim, "Prediction of type 2 diabetes based on machine learning algorithm," *International Journal of Environmental Research and Public Health*, vol. 18, no. 6, Mar. 2021, doi: 10.3390/ijerph18063317.

[14] M. Z. Al Yousef, A. F. Yasky, R. Al Shammari, and M. S. Ferwana, "Early prediction of diabetes by applying data mining techniques: a retrospective cohort study," *Medicine*, vol. 101, no. 29, Jul. 2022, doi: 10.1097/MD.0000000000029588.

[15] H. El-Sofany, S. A. El-Seoud, O. H. Karam, Y. M. A. El-Latif, and I. A. T. F. T. Eddin, "A proposed technique using machine learning for the prediction of diabetes disease through a mobile app," *International Journal of Intelligent Systems*, pp. 1–13, 2024, doi: 10.1155/2024/6688934.
[16] R. Hasan, V. Dattana, S. Mahmood, and S. Hussain, "Towards transparent diabetes prediction: combining AutoML and explainable AI for improved clinical insights," *Information*, vol. 16, no. 1, Dec. 2024, doi: 10.3390/info16010007.
[17] Y. Zhao, J. K. Chaw, M. C. Ang, M. M. Daud, and L. Liu, "A diabetes prediction model with visualized explainable artificial intelligence (XAI) technology," in *Advances in Visual Informatics: 8th International Visual Informatics Conference (IVIC 2023), Selangor, Malaysia*, 2024, pp. 648–661, doi: 10.1007/978-981-99-7339-2_52.
[18] P. Netayawijit, W. Chansanam, and K. S. In, "Interpretable machine learning framework for diabetes prediction integrating SMOTE balancing with SHAP explainability," *Healthcare*, vol. 13, no. 20, Oct. 2025, doi: 10.3390/healthcare13202588.
[19] R. Ganguly and D. Singh, "Explainable artificial intelligence (XAI) for the prediction of diabetes management: an ensemble approach," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 7, pp. 158–163, 2023, doi: 10.14569/IJACSA.2023.0140717.
[20] X.-E. Gao, J.-G. Hu, B. Chen, Y.-M. Wang, and S.-B. Zhou, "Causal discovery approach with reinforcement learning for risk factors of type II diabetes mellitus," *BMC Bioinformatics*, vol. 24, no. 1, Jul. 2023, doi: 10.1186/s12859-023-05405-x.
[21] A. Descarpentrie *et al.*, "Social and behavioral factors associated with diabetes in Southern California vs the US," *JAMA Network Open*, vol. 8, no. 10, Oct. 2025, doi: 10.1001/jamanetworkopen.2025.38377.
[22] S.-Y. Lee *et al.*, "Integrating social determinants of health in machine learning-driven decision support for diabetes case management: protocol for a sequential mixed methods study," *JMIR Research Protocols*, vol. 13, 2024, doi: 10.2196/56049.
[23] Z. Javed, H. Kundi, R. Chang, A. Titus, and H. Arshad, "Polysocial risk scores: implications for cardiovascular disease risk assessment and management," *Current Atherosclerosis Reports*, vol. 25, no. 12, pp. 1059–1068, 2023, doi: 10.1007/s11883-023-01173-4.
[24] R. Jain and N. K. Tripathi, "Exploring diabetes prediction, gender, and age variability for effective diabetes management," in *Smart Healthcare, Clinical Diagnostics, and Bioprinting Solutions for Modern Medicine*, Palmdale, United States: IGI Global Scientific Publishing, 2025, pp. 151–164, doi: 10.4018/979-8-3373-0659-9.ch008.
[25] G. Husain *et al.*, "SMOTE vs. SMOTEENN: a study on the performance of resampling algorithms for addressing class imbalance in regression models," *Algorithms*, vol. 18, no. 1, Jan. 2025, doi: 10.3390/a18010037.
[26] J. Tuomilehto, M. Uusitupa, E. W. Gregg, and J. Lindström, "Type 2 diabetes prevention programs—from proof-of-concept trials to national intervention and beyond," *Journal of Clinical Medicine*, vol. 12, no. 5, Feb. 2023, doi: 10.3390/jcm12051876.

BIOGRAPHIES OF AUTHORS

Mohammad Saeed Al Ghamdi received the B.Sc. degree in Information Technology from Arab Open University, Jeddah, Saudi Arabia, in 2017. He is currently a M.Sc. student with the Faculty of Computing and Information Systems, King Abdulaziz University, Jeddah, Saudi Arabia. His main research interests include ML and medical data analysis. He can be contacted at email: mmathkoralghamdi@stu.kau.edu.sa.

Dr. Alaa Khadidos received the B.Sc. degree from King Abdulaziz University, Jeddah, Saudi Arabia, in 2006, the M.Sc. degree from the University of Birmingham, Birmingham, United Kingdom, in 2011, and the Ph.D. degree from the University of Warwick, Coventry, United Kingdom, in 2017, all in Computer Science. He is currently an assistant professor with the Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia. His main research interests include computer vision, ML, optimization, and medical image analysis. He can be contacted at email: aokhadidos@kau.edu.sa.

Dr. Adil Khadidos received the B.Sc. degree in Computer Science from King Abdulaziz University, Jeddah, Saudi Arabia, in 2006, the M.Sc. degree in Internet Software Systems from the University of Birmingham, Birmingham, United Kingdom, in 2011, and the Ph.D. degree in Computer Science from the University of Southampton, Southampton, United Kingdom, in 2017. He is currently an assistant professor at the Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia. His main research interests include computer swarm robotics, insect behavior, ML, self-distributed systems, and embedded systems. He can be contacted at email: akhadidos@kau.edu.sa.