Cite this article: Hanmanop, S., Jongsiri, T., Waiprasit, K., & Pechprasarn, S. (2026). A data-driven framework for diabetes prediction: Machine learning-based comparison of invasive and non-invasive screening. *Journal of Current Science and Technology*, *16*(1), Article 159. https://doi.org/10.59796/jcst.V16N1.2026.159

Journal of Current Science and Technology

Journal homepage: https://jcst.rsu.ac.th

A Data-Driven Framework for Diabetes Prediction: Machine Learning-Based Comparison of Invasive and Non-Invasive Screening

Sasipatcha Hanmanop[1], Tatpol Jongsiri[1], Kittitat Waiprasit[1], and Suejit Pechprasarn[1, 2, *]

[1]College of Biomedical Engineering, Rangsit University, Pathum Thani 12000, Thailand
[2]Center of Excellence in AI and Supercomputing, Rangsit University, Pathum Thani 12000, Thailand

*Corresponding author: suejit.p@rsu.ac.th

Received 30 April 2025; Revised 4 August 2025; Accepted 15 September 2025; Published online 25 December 2025

Abstract

This study evaluated the performance of 24 models for diabetes prediction by using eight predictors: sex, heart disease, hypertension, smoking history, BMI (Body Mass Index), HbA1c level (Hemoglobin A1c), and blood glucose level obtained from an open data source, Kaggle. Data preparation involved curating and cleaning to ensure unbiased training and a balanced dataset before applying the dataset to machine learning training. The research examined data splitting ratios at 70/30, 80/20, and 90/10. The prediction task focused on the diabetes category: 0 (non-diabetes) and 1 (diabetes). The performance parameters indicated that the Ensemble Boosted Trees model, particularly with a 70/30 data splitting ratio, achieved the highest accuracy of 91.45%, precision of 91.29%, recall of 91.65%, and F1-score of 91.37%. Feature selection, including Chi-Square (χ^2) ANOVA, Kruskal-Wallis, and principal component analysis have been applied to reduce the complexity and dimensionality of the model, and it was found that the following parameters were significant for diabetes diagnosis: (1) HbA1c, (2) blood glucose, (3) BMI, and (4) age. The first two parameters are crucial for medical practitioners to determine whether a patient has diabetes; however, they are invasive and can only be collected from blood test results. Here, we also discuss the accuracy of the machine learning model in predicting diabetes without invasive predictors, namely, blood glucose and HbA1c. Our simplified model using age and BMI still yielded a reasonable accuracy of 74.65%, demonstrating the feasibility of non-blood test and non-invasive screening, especially in resource-limited settings, where age and BMI are key non-blood test predictors.

Keywords: *artificial intelligence; diabetes; feature selection method; dimension reduction method; machine learning; non-blood test diabetes prediction*

1. Introduction

Diabetes mellitus is a chronic metabolic disorder characterized by impaired glucose regulation resulting from insufficient insulin production or ineffective insulin utilization (Alam et al., 2021; Sarkar et al., 2019). This results in glucose accumulation in the bloodstream, causing a metabolic imbalance that, if unsolved, can result in several health complications, including cardiovascular disease, neuropathy, vision problems, and, ultimately, organ damage (Calibo, 2024; Prabhakar et al., 2024).

Diabetes presents a significant global health challenge, affecting an estimated 10.5% of the world's population, with an estimated 536.6 and 783.2 million adults living with the condition (Boadu et al., 2024). Several forms of diabetes exist, each with distinct challenges and implications for glucose metabolism, including Type 1, Type 2, gestational diabetes, secondary diabetes, maturity-onset diabetes of the young, and latent autoimmune diabetes in adults (Skyler et al., 2017; Buzzetti et al., 2017). Clinical manifestations often develop insidiously and include polydipsia, polyuria, unexplained weight loss, fatigue, and visual disturbances (Jalilian et al., 2023; Looareesuwan et al., 2023).

Early diagnosis of diabetes is crucial, as patients often remain asymptomatic in the beginning stages (Carmichael et al., 2021; Pippitt et al., 2016). According to clinical practice guidelines, screening is advised based on risk, at any age for overweight or obese individuals, and beginning at 35 for individuals not in those categories. (Tiwari & Aw, 2024; Davidson et al., 2024). Traditional diagnostic methods, including fasting plasma glucose and HbA1c testing, have inherent limitations such as cost, invasiveness, and accessibility barriers, frequently resulting in delayed detection until complications emerge (Bergman et al., 2020; Zhang et al., 2023).

Artificial intelligence (AI), especially machine learning, offers a paradigm shift in diabetes diagnosis by analyzing complex datasets, including patient records and blood glucose measurements, to identify subtle patterns representative of early-stage diabetes more efficiently and accurately than old-fashioned approaches (Dagliati et al., 2018; Poorani et al., 2025 ; Anupongongarch et al., 2022). However, challenges remain, such as selecting relevant features and translating predictions into clinical action. This study addresses these challenges by systematically evaluating multiple machine learning models and feature selection strategies to develop an accurate, non-invasive approach for early diabetes screening.

Ghosh et al. (2021) examine four machine learning-based classifiers, achieving an accuracy of 99.35% for the Random Forest classifier, which included interventional methods such as the insulin test and glucose level parameters. Similarly, Dritsas & Trigka (2022) developed multiple models and evaluated their performance using 10-fold cross-validation with the Synthetic Minority Over-sampling Technique (SMOTE) and data splitting. Their best-performing model, which excluded invasive features, achieved a top accuracy of 99.22%. Qin et al. (2022) demonstrated the effectiveness of CATBoost, XGBoost, Logistic Regression, Random Forest, and Support Vector Machine classifiers, achieving 82.1% accuracy and an Area Under the Receiver Operating Characteristic Curve (AUC) of 0.83. These classifiers were developed using parameters that included blood test results and primarily focused on lifestyle-type parameters. While most existing studies rely heavily on blood test parameters (Chapakiya et al., 2025; Khanam & Foo, 2021; Rani, 2020), there is limited research on accurate diabetes risk prediction using only non-invasive, readily available measurements. Such an approach could enhance accessibility, reduce costs, and enable preliminary screening in resource-constrained settings without requiring invasive blood tests. Previous research and statistical evaluations have consistently highlighted four key predictors as most influential in diabetes diagnosis: Glycated Hemoglobin (HbA1c), blood glucose level, Body Mass Index (BMI), and age. Notably, while HbA1c and blood glucose are derived from invasive blood tests, BMI and age are non-invasive and readily accessible (Liu et al., 2025; Sinsophonphap & Thavornsawadi, 2022). This distinction underpins the potential for developing simplified, accessible screening tools using non-blood-based indicators.

In this study, we utilize the open-source dataset from Mohammed Mustafa on Kaggle (Mustafa, 2023), which contains 100,000 patient records with eight clinical predictors. Our method customizes refining diagnostic indicators to improve accuracy and interpretability by reducing blood test predictors to develop a non-invasive diabetes diagnosis approach and compare the performance of the two approaches. We evaluated 24 machine learning algorithms to identify optimal predictive models and determine the most influential features for diabetes screening. Feature selection methods are essential for reducing model complexity, enhancing predictor performance, and expanding previous research (Pechprasarn et al., 2025) on ML algorithm parameters for diabetes prediction.

2. Objectives

1. To improve diabetes prediction and diagnosis by developing and evaluating machine learning models that accurately classify diabetic and non-diabetic individuals.

2. To analyze the impact of different training to test data ratios on model performance.

3. To develop a self-assessment method for early diabetes detection through non-invasive predictors that do not require blood tests, which can improve accessibility while reducing computational complexity.

3. Materials and Methods

We collected and curated the data for this research to ensure that the diabetes and non-diabetes datasets had equal samples. The data rows were separated into training and test datasets at 70/30, 80/20, and 90/10 ratios. All 24 models used to train the data are available in MATLAB R2024a and are included here. The flow of this research is depicted in Figure 1.

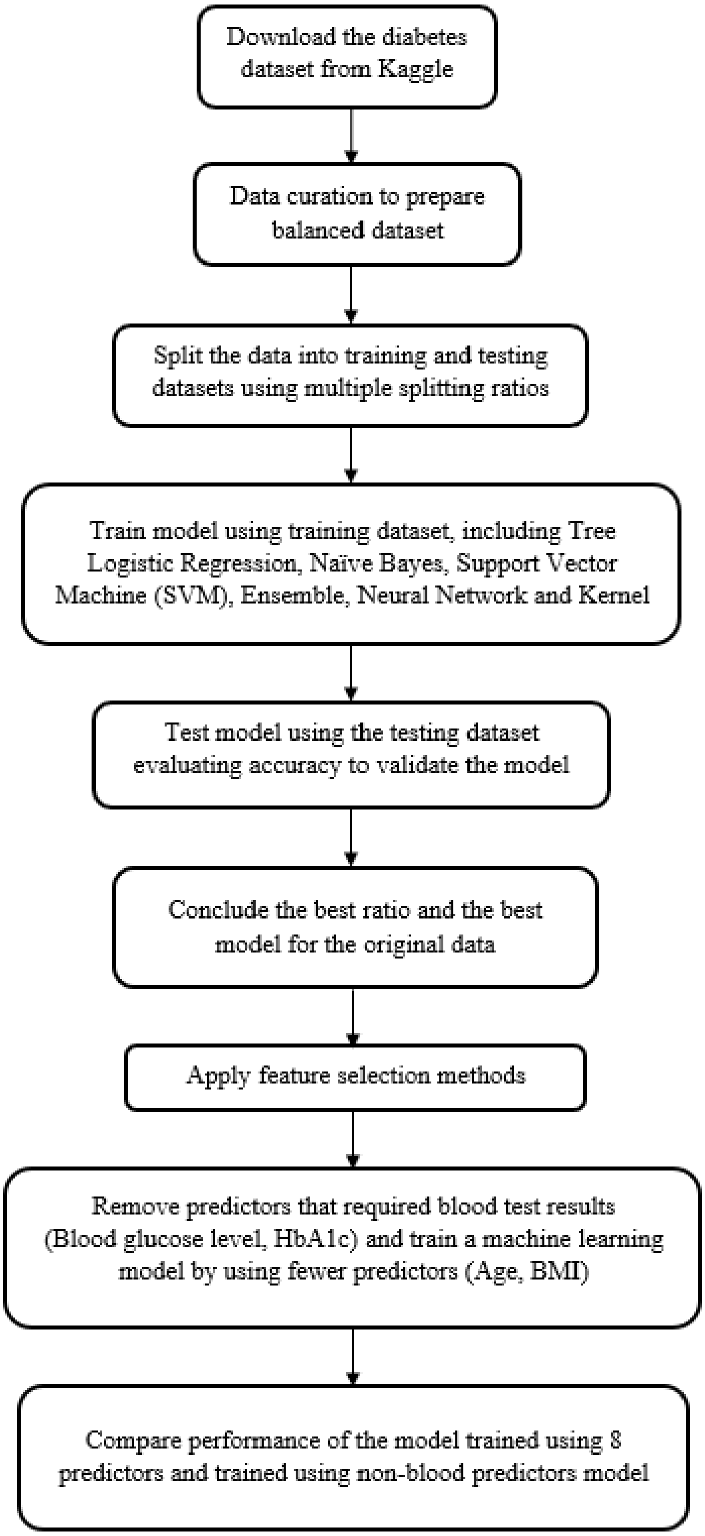

Download the diabetes dataset from Kaggle
Data curation to prepare balanced dataset
Split the data into training and testing datasets using multiple splitting ratios
Train model using training dataset, including Tree Logistic Regression, Naïve Bayes, Support Vector Machine (SVM), Ensemble, Neural Network and Kernel
Test model using the testing dataset evaluating accuracy to validate the model
Conclude the best ratio and the best model for the original data
Apply feature selection methods
Remove predictors that required blood test results (Blood glucose level, HbA1c) and train a machine learning model by using fewer predictors (Age, BMI)
Compare performance of the model trained using 8 predictors and trained using non-blood predictors model

Figure 1 Research process flow for data preparation and machine learning model evaluation. Modified from “Predicting Parkinson’s Disease Severity Using Telemonitoring Data and Machine Learning Models: A Principal Component Analysis-Based Approach for Remote Healthcare Services During the COVID-19 Pandemic” (Pechprasarn et al., 2023)

Table 1 Description of clinical predictors, data types, and value ranges

Variable	Predictor/Label in Machine Learning	Category	Values/Value Range
Diabetes	Label	Nominal	0 = Non-diabetes 1 = Diabetes
Heart Disease	Predictor	Nominal	0 = Non-Heart Disease 1 = Heart disease
Hypertension	Predictor	Nominal	0 = There is no continuous increase in blood pressure levels. 1 = There is a continuous increase in blood pressure levels.
Smoking History	Predictor	Nominal	No Info = No record of smoking history. Never = Has not smoked regularly Former = Previously smoked but no longer does. Current = Actively smoking. Ever = often, smoking at least once per day.
BMI (Body Mass Index)	Predictor	Numerical	In the kg/m^2 unit
HbA1c level (Hemoglobin A1c)	Predictor	Numerical	In mg% unit
Blood Glucose level	Predictor	Numerical	In the mg/dL unit
Sex	Predictor	Nominal	Male and Female
Age	Predictor	Nominal	Patients' age in year

3.1 Dataset Details

The dataset for estimating diabetes in this paper was obtained from Mohammed Mustafa's data on the Kaggle website (Mustafa, 2023); the data file is in Excel format with a .csv extension. This dataset has 100,000 patient data points and eight columns of predictors for training supervised ML models. Variables used are sex, age, hypertension, heart disease, smoking history, BMI, Glycated hemoglobin level (HbA1c), and blood glucose level, as shown in Table 1.

3.2 Data Curation

The data was processed by curating and cleaning the diabetic dataset to ensure unbiased training. The dataset that was obtained has a significant imbalance. The diabetes dataset contains 8,500 patients, while the non-diabetes dataset has 91,500 patients. Imbalanced data can impact the machine learning model, causing it to predict "non-diabetes" more frequently than "diabetes", resulting in unfair predictions. To prevent prediction errors caused by this imbalance, the dataset should be adjusted to achieve a balanced distribution. This balanced dataset was created by randomly eliminating 83,000 non-diabetes entries, both classes have 8,500 cases each from this method.

3.3 Dataset for Training and Testing

This dataset should be divided into two files for data curation: model training and testing. The ratio of training and testing datasets is determined by comparing the ratios of 70/30, 80/20, and 90/10 for training and testing, as shown in Table 2.

Table 2 Training and Testing Dataset Splits by Ratio

Ratio	Train	Test	Total
70/30	11,900	5,100	17,000
80/20	13,600	3,400	17,000
90/10	15,300	1,700	17,000

3.4 Machine Learning Training and Testing

After splitting the data at different ratios, we compared the accuracy of each ratio. We trained and tested these datasets using 24 models in MATLAB R2024a, as listed in Table 3. For the test dataset, we used the same performance matrix as the training dataset for easy comparison, including a K-fold cross-validation (K-fold=5) to calculate performance metrics, including precision, recall, accuracy, and F1-score, using the training dataset as demonstrated in equations (1) - (4) as follows:

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

$$\text{Precision} = \frac{\text{TP}}{\text{TP + FP}} \quad (2)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP + FN}} \quad (3)$$

$$\text{F1-score} = \frac{\text{2 x Precision x Recall}}{\text{Precision + Recall}} \quad (4)$$

From this equation, True Positive (TP) refers to the correct prediction of patients with diabetes, True Negative (TN) refers to the correct prediction of patients with non-diabetes, and A False Positive (FP) occurs when non-diabetes is misclassified as diabetes, and a False Negative (FN) occurs when diabetes is misclassified as non-diabetes. The results obtained from the calculations will indicate the effectiveness of the model used for training and testing, ensuring that our model can be improved and developed to enhance diabetes diagnosis in medical practice.

3.5 Feature Selection and Simplified Model

Feature selection algorithms, including Chi-Square (χ^2) ANOVA and the Kruskal-Wallis algorithms identify the crucial factors that reduce the number of predictors used to train machine learning models. This process involves adding predictors one by one and comparing the accuracy results to determine how many predictors are necessary for this machine learning model, as well as identifying those that are less important, thereby reducing the processing complexity. These have been widely used for dimensionality reduction. In this research, the dimension reduction method was applied using MATLAB to identify the importance and relevance of each predictor (blood glucose levels, HbA1c, heart disease, hypertension, smoking history, BMI, sex) to find the most important predictors, excluding blood test predictors (HbA1c, blood glucose levels), for developing a non-invasive diabetes diagnosis approach Less important predictors were removed, and the remaining ones were analyzed to assess whether simple, non-blood-based indicators could be used effectively for early diabetes screening (Pechprasarn et al., 2023).

4. Results

This paper explores how AI and machine learning methods can be improved to achieve the best performance in predicting diabetes. This project will compare the accuracy of different ML models and ratios of training and testing datasets to determine the most effective approach.

4.1 Training of Classification Models

During the model training phase, this study utilized dataset splits of 70/30, 80/20, and 90/10 for training and testing. All machine learning models were trained using the training dataset, and their performance was compared and validated based on key evaluation metrics: precision, recall, F1-score, and accuracy using 5-fold cross-validation, as presented in Table 4.

Table 3 24 Models available in MATLAB R2024a

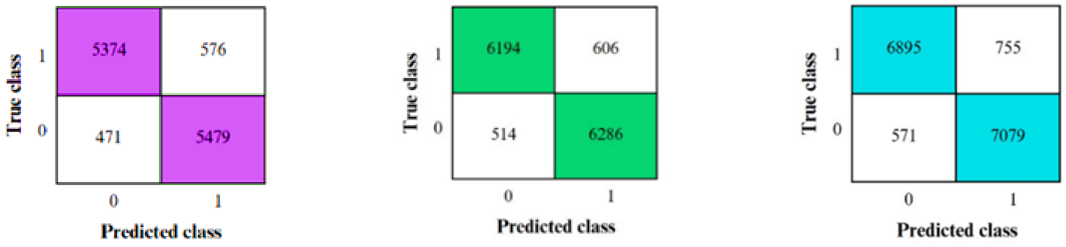
Model	Details	Model	Details
Tree	Fine Tree	SVM	Linear SVM
	Medium Tree		Quadratic SVM
	Coarse Tree		Cubic SVM
Ensemble	Boosted Trees		Fine Gaussian SVM
	Bagged Trees		Medium Gaussian SVM
	RUS Boosted Tree		Coarse Gaussian SVM
Kernel	SVM Kernel	Neural Network	Narrow Neural Network
	Logistic Regression Kernel		Medium Neural Network
Naïve Bayes	Gaussian Naïve Bayes		Wide Neural Network
	Kernel Naïve Bayes		Bilayer Neural Network
Binary GLM Logistic Regression	Binary GLM Logistic Regression		Tri-layered Neural Network
Efficient Logistic Regression	Efficient Logistic Regression	Efficient Linear SVM	Efficient Linear SVM

Table 4 Cross-validation results of 24 models across different data split ratios

Model performance computed using 5-fold cross-validation with a dataset split ratio of 70/30					
Model	**Details**	**Accuracy**	**Precision**	**Recall**	**F1-score**
Tree	Fine Tree	89.72%	89.19%	90.40%	89.46%
	Medium Tree	90.10%	89.49%	90.87%	89.79%
	Coarse Tree	85.05%	79.49%	94.47%	82.18%
Binary GLM Logistic Regression	Binary GLM Logistic Regression	88.24%	88.36%	88.07%	88.30%
Efficient Logistic Regression	Efficient Logistic Regression	88.39%	88.52%	88.22%	88.45%
Efficient Linear SVM	Efficient Linear SVM	87.74%	88.47%	86.79%	88.10%
Naïve Bayes	Gaussian Naïve Bayes	83.99%	88.43%	78.22%	86.15%
	Kernel Naïve Bayes	90.36%	90.46%	90.24%	90.41%
SVM	Linear SVM	88.46%	88.34%	88.62%	88.40%
	Quadratic SVM	88.61%	88.46%	88.79%	88.53%
	Cubic SVM	89.95%	89.19%	90.92%	89.57%
	Fine Gaussian SVM	87.85%	85.67%	90.91%	86.74%
	Medium Gaussian SVM	89.47%	88.71%	90.45%	89.09%
	Coarse Gaussian SVM	88.61%	88.43%	88.86%	88.52%
Ensemble	Boosted Trees*	91.20%	90.49%	92.08%	90.84%
	Bagged Trees	89.99%	89.73%	90.32%	89.86%
	RUS Boosted Tree	90.10%	89.49%	90.87%	89.79%
Neural Network	Narrow Neural Network	90.48%	90.12%	90.92%	90.30%
	Medium Neural Network	89.81%	89.51%	90.18%	89.66%
	Wide Neural Network	87.82%	88.12%	87.43%	87.97%
	Bilayer Neural Network	90.08%	89.28%	91.09%	89.67%
	Tri-layered Neural Network	89.94%	89.13%	90.97%	89.53%
Kernel	SVM Kernel	88.82%	87.83%	90.13%	88.32%
	Logistic Regression Kernel	87.98%	86.94%	89.39%	87.46%
Model performance computed using 5-fold cross-validation with a dataset split ratio of 80/20					
Model	**Details**	**Accuracy**	**Precision**	**Recall**	**F1-score**
Tree	Fine Tree	89.98%	89.96%	90.00%	89.97%
	Medium Tree	90.62%	90.25%	91.07%	90.43%
	Coarse Tree	86.21%	81.87%	93.00%	83.98%
Binary GLM Logistic Regression	Binary GLM Logistic Regression	88.51%	88.49%	88.54%	88.50%
Efficient Logistic Regression	Efficient Logistic Regression	88.48%	88.60%	88.32%	88.54%
Efficient Linear SVM	Efficient Linear SVM	88.64%	88.54%	88.76%	88.59%
Naïve Bayes	Gaussian Naïve Bayes	84.19%	88.70%	78.37%	86.39%
	Kernel Naïve Bayes	90.42%	90.45%	90.38%	90.43%
SVM	Linear SVM	88.67%	88.38%	89.04%	88.52%
	Quadratic SVM	88.82%	88.21%	89.63%	88.51%
	Cubic SVM	90.44%	89.64%	91.46%	90.04%
	Fine Gaussian SVM	88.61%	86.48%	91.53%	87.53%
	Medium Gaussian SVM	89.74%	88.94%	90.75%	89.34%
	Coarse Gaussian SVM	88.74%	88.36%	89.24%	88.55%
Ensemble	Boosted Trees	91.76%	91.21%	92.44%	91.49%
	Bagged Trees	90.50%	90.30%	90.75%	90.40%
	RUS Boosted Tree	90.60%	90.21%	91.07%	90.40%
Neural Network	Narrow Neural Network	90.80%	90.52%	91.15%	90.66%
	Medium Neural Network	90.33%	90.35%	90.31%	90.34%
	Wide Neural Network	88.62%	88.57%	88.68%	88.59%
	Bilayer Neural Network	90.68%	90.13%	91.35%	90.40%
	Tri-layered Neural Network	90.35%	90.07%	90.71%	90.21%
Kernel	SVM Kernel	88.99%	87.92%	90.41%	88.45%
	Logistic Regression Kernel	88.03%	86.85%	89.63%	87.43%

Table 4 Cont.

Model performance computed using 5-fold cross-validation with a dataset split ratio of 90/10					
Model	**Details**	**Accuracy**	**Precision**	**Recall**	**F1-score**
Tree	Fine Tree	90.25%	89.46%	91.27%	89.85%
	Medium Tree	90.46%	89.86%	91.20%	90.16%
	Coarse Tree	84.53%	78.11%	95.95%	81.19%
Binary GLM Logistic Regression	Binary GLM Logistic Regression	88.39%	88.46%	88.29%	88.42%
Efficient Logistic Regression	Efficient Logistic Regression	81.19%	88.55%	88.29%	84.71%
Efficient Linear SVM	Efficient Linear SVM	81.12%	88.50%	88.21%	84.65%
Naïve Bayes	Gaussian Naïve Bayes	83.87%	88.57%	78.09%	86.16%
	Kernel Naïve Bayes	90.34%	90.39%	90.27%	90.37%
SVM	Linear SVM	88.51%	88.42%	88.63%	88.46%
	Quadratic SVM	88.88%	88.48%	89.39%	88.68%
	Cubic SVM	90.39%	89.95%	90.93%	90.17%
	Fine Gaussian SVM	88.44%	86.71%	90.78%	87.57%
	Medium Gaussian SVM	89.65%	89.09%	90.37%	89.37%
	Coarse Gaussian SVM	88.63%	88.45%	88.86%	88.54%
Ensemble	Boosted Trees	91.33%	90.36%	92.54%	90.85%
	Bagged Trees	90.16%	89.94%	90.44%	90.05%
	RUS Boosted Tree	90.46%	89.88%	91.20%	90.17%
Neural Network	Narrow Neural Network	90.66%	90.40%	90.98%	90.53%
	Medium Neural Network	90.14%	90.12%	90.16%	90.13%
	Wide Neural Network	88.59%	88.87%	88.22%	88.73%
	Bilayer Neural Network	90.41%	89.91%	91.05%	90.16%
	Tri-layered Neural Network	90.52%	90.05%	91.10%	90.28%
Kernel	SVM Kernel	88.90%	88.04%	90.03%	88.46%
	Logistic Regression Kernel	88.69%	87.88%	89.75%	88.28%



(a) (b) (c)

Figure 2 Confusion matrices of the trained Ensemble Boosted Trees model at different train-test split ratios: (a) 70/30, (b) 80/20, (c) 90/10

Among the 24 trained models, the Ensemble Boosted Trees model for diabetes prediction and diagnosis produced the best results. The three train-test split ratios were evaluated: 70/30, 80/20, and 90/10. With a 70/30 split, the Ensemble Boosted Trees model achieved an accuracy of 91.20%, a precision of 90.49%, a recall of 92.08%, and an F1-score of 90.84%. With an 80/20 split, the model achieved an accuracy of 91.76%, a precision of 91.21%, a recall of 92.44%, and an F1-score of 91.49%. For the 90/10 split, the model achieved an accuracy of 91.33%, a precision of 90.36%, a recall of 92.54%, and an F1-score of 90.85%. The confusion matrices of the three splitting ratios are shown in Figure 2.

4.2 Performance Evaluation Using the Test Dataset

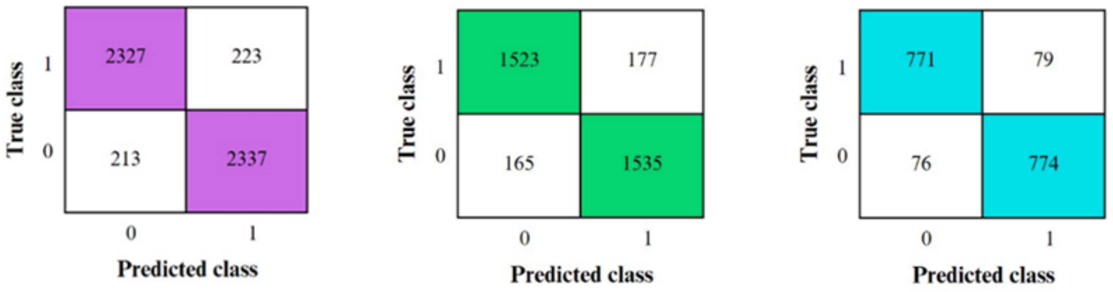
After training the models, the trained models were tested using a separate unseen dataset reserved for evaluation. The testing dataset was separated using the same procedure as the training dataset. Model performance for the test dataset was assessed using the same performance metrics for a direct comparison to the validation performance. The detailed results for each split ratio are presented in Table 5.

Table 5 Test dataset performance of 24 models across different data split ratios

Model performance was evaluated using the unseen test dataset with a dataset split ratio of 70/30					
Model	**Details**	**Accuracy**	**Precision**	**Recall**	**F1-score**
Tree	Fine Tree	90.18%	89.24%	91.37%	89.70%
	Medium Tree	90.43%	90.53%	90.31%	90.48%
	Coarse Tree	85.31%	79.55%	95.06%	82.33%
Binary GLM Logistic Regression	Binary GLM Logistic Regression	88.35%	88.60%	88.04%	88.47%
Efficient Logistic Regression	Efficient Logistic Regression	88.51%	88.84%	88.08%	88.68%
Efficient Linear SVM	Efficient Linear SVM	88.57%	88.64%	88.47%	88.61%
Naïve Bayes	Gaussian Naïve Bayes	84.02%	88.68%	78.00%	86.28%
	Kernel Naïve Bayes	90.20%	90.13%	90.27%	90.16%
SVM	Linear SVM	88.47%	88.41%	88.55%	88.44%
	Quadratic SVM	88.69%	88.61%	88.78%	88.65%
	Cubic SVM	90.37%	89.40%	91.61%	89.88%
	Fine Gaussian SVM	88.49%	86.66%	90.98%	87.57%
	Medium Gaussian SVM	89.61%	89.09%	90.27%	89.35%
	Coarse Gaussian SVM	88.59%	88.56%	88.63%	88.57%
Ensemble	Boosted Trees	91.45%	91.29%	91.65%	91.37%
	Bagged Trees	90.59%	90.68%	90.47%	90.64%
	RUS Boosted Tree	90.43%	90.53%	90.31%	90.48%
Neural Network	Narrow Neural Network	90.55%	89.98%	91.25%	90.27%
	Medium Neural Network	90.39%	89.68%	91.29%	90.03%
	Wide Neural Network	89.55%	89.97%	89.02%	89.76%
	Bilayer Neural Network	90.37%	88.86%	92.31%	89.61%
	Tri-layered Neural Network	90.61%	90.09%	91.25%	90.35%
Kernel	SVM Kernel	88.94%	88.10%	90.04%	88.52%
	Logistic Regression Kernel	88.59%	87.90%	89.49%	88.25%
Model performance evaluated using the unseen test dataset with a dataset split ratio of 80/20					
Model	**Details**	**Accuracy**	**Precision**	**Recall**	**F1-score**
Tree	Fine Tree	88.71%	88.52%	88.94%	88.62%
	Medium Tree	88.56%	89.13%	87.82%	88.85%
	Coarse Tree	83.76%	78.05%	93.94%	80.81%
Binary GLM Logistic Regression	Binary GLM Logistic Regression	87.76%	87.81%	87.71%	87.79%
Efficient Logistic Regression	Efficient Logistic Regression	88.00%	88.00%	88.00%	88.00%
Efficient Linear SVM	Efficient Linear SVM	87.97%	88.31%	87.53%	88.14%
Naïve Bayes	Gaussian Naïve Bayes	83.35%	87.35%	78.00%	85.31%
	Kernel Naïve Bayes	89.44%	89.10%	89.88%	89.27%
SVM	Linear SVM	87.91%	87.76%	88.12%	87.83%
	Quadratic SVM	88.18%	87.82%	88.65%	88.00%
	Cubic SVM	89.09%	88.34%	90.06%	88.71%
	Fine Gaussian SVM	87.53%	85.29%	90.71%	86.39%
	Medium Gaussian SVM	88.88%	88.30%	89.65%	88.59%
	Coarse Gaussian SVM	87.79%	87.55%	88.12%	87.67%
Ensemble	Boosted Trees	89.94%	89.66%	90.29%	89.80%
	Bagged Trees	89.50%	89.29%	89.76%	89.40%
	RUS Boosted Tree	88.56%	89.13%	87.82%	88.85%
Neural Network	Narrow Neural Network	89.74%	89.30%	90.29%	89.52%
	Medium Neural Network	89.32%	89.12%	89.59%	89.22%
	Wide Neural Network	88.35%	88.95%	87.59%	88.65%
	Bilayer Neural Network	89.29%	89.34%	89.24%	89.32%
	Tri-layered Neural Network	89.53%	88.75%	90.53%	89.14%
Kernel	SVM Kernel	87.56%	86.26%	89.35%	86.90%
	Logistic Regression Kernel	86.71%	85.58%	88.29%	86.14%

Table 5 Cont.

Model performance evaluated using the unseen test dataset with a dataset split ratio of 90/10					
Model	**Details**	**Accuracy**	**Precision**	**Recall**	**F1-score**
Tree	Fine Tree	90.60%	89.82%	90.35%	89.94%
	Medium Tree	89.47%	89.24%	89.76%	89.36%
	Coarse Tree	83.71%	77.31%	95.41%	80.38%
Binary GLM Logistic Regression	Binary GLM Logistic Regression	87.76%	88.03%	87.41%	87.90%
Efficient Logistic Regression	Efficient Logistic Regression	87.88%	88.33%	87.29%	88.11%
Efficient Linear SVM	Efficient Linear SVM	87.00%	88.40%	85.18%	87.69%
Naïve Bayes	Gaussian Naïve Bayes	82.88%	88.03%	76.12%	85.38%
	Kernel Naïve Bayes	89.59%	89.73%	89.41%	89.66%
SVM	Linear SVM	87.65%	87.91%	87.29%	87.78%
	Quadratic SVM	87.94%	87.99%	87.88%	87.96%
	Cubic SVM	90.39%	89.95%	90.93%	90.17%
	Fine Gaussian SVM	87.88%	86.51%	89.76%	87.19%
	Medium Gaussian SVM	89.00%	89.05%	88.94%	89.02%
	Coarse Gaussian SVM	87.88%	88.06%	87.65%	87.97%
Ensemble	Boosted Trees	90.88%	90.74%	91.06%	90.81%
	Bagged Trees	89.88%	89.60%	90.24%	89.74%
	RUS Boosted Tree	89.47%	89.24%	89.76%	89.36%
Neural Network	Narrow Neural Network	90.18%	90.41%	89.88%	90.30%
	Medium Neural Network	90.41%	90.84%	89.88%	90.63%
	Wide Neural Network	88.94%	89.50%	88.24%	89.22%
	Bilayer Neural Network	90.41%	90.46%	90.35%	90.44%
	Tri-layered Neural Network	89.94%	90.18%	89.65%	90.06%
Kernel	SVM Kernel	87.06%	86.63%	87.65%	86.84%
	Logistic Regression Kernel	87.24%	86.51%	88.24%	86.87%



(a) (b) (c)

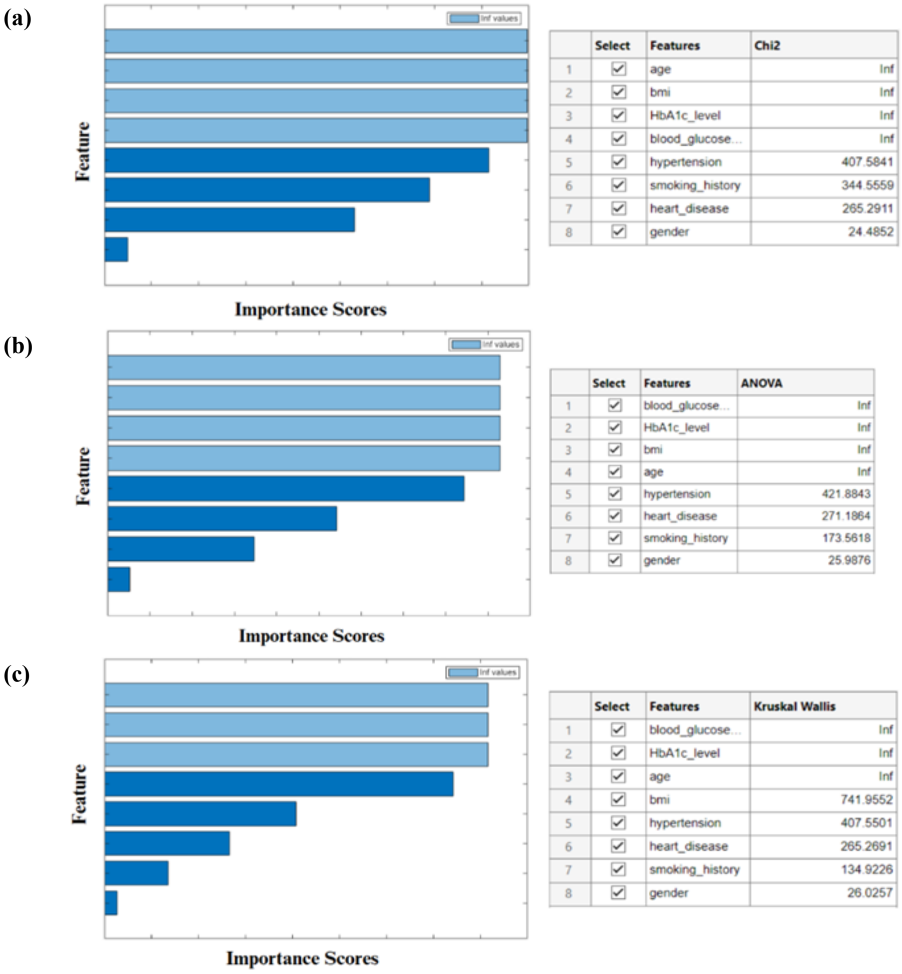
Figure 3 Confusion matrices of the Ensemble Boosted Trees model on the test dataset for three split ratios: (a) 70/30, (b) 80/20, (c) 90/10

For the 24 trained models evaluated on the test dataset, the Ensemble Boosted Trees model for diabetes prediction and diagnosis produced the best results. Three train-test split ratios were evaluated: 70/30, 80/20, and 90/10. With a 70/30 split, the Ensemble Boosted Trees model achieved an accuracy of 91.45%, a precision of 91.29%, a recall of 91.65%, and an F1-score of 91.37%. With an 80/20 split, the model achieved an accuracy of 89.94%, a precision of 89.66%, a recall of 90.29%, and an F1-score of 89.80%. For the 90/10 split, the model achieved an accuracy of 90.88%, a precision of 90.74%, a recall of 91.06%, and an F1-score of 90.81%. Therefore, it can be concluded that the dataset size was sufficiently large for all the ratios, and there was no significant difference among the three cases. The performance of the 5-fold cross-validation in Table 4 and the test performance in Table 5 were well within 1.8% for the ensemble-boosted tree models, indicating an optimal fit for the models. The confusion matrices of the three splitting ratios for the test cases are shown in Figure 3.

4.3 Feature Selection and Dimension Reduction

Here, three statistical algorithms, including Chi-Square (χ^2) ANOVA and the Kruskal-Wallis algorithms were employed to rank the importance of each feature for the eight predictors (including HbA1c level and blood glucose level). The statistical values of the three ranking algorithms are shown in Figure 4. We found that four predictors, including HbA1c, blood glucose levels, age, and BMI, have a statistical value of infinity across all three algorithms. Even though BMI does not have an infinite value in the Kruskal-Wallis algorithm, the value is still considered high.

HbA1c and blood glucose levels are obtained from blood tests to measure blood sugar levels, two of the four most essential predictors in determining diabetes in medical practice. In other words, if blood test results are available, these values can indicate whether a person has diabetes. The Ensemble Boosted Trees model was trained by adding one predictor at a time, as shown in Table 6. The model trained using three predictors achieved an accuracy of 87.96%, which was 5% higher than the model trained using two blood-related parameters, indicating that the BMI factor plays a crucial role in diabetes prediction. Adding the age parameter improved the accuracy performance further by around 2%, achieving 90.79% accuracy. The rest of the parameters, after four predictors, did not show significant improvement. Based on these findings, this research concludes that only four predictors are necessary for diabetes prediction, effectively reducing the computational complexity of machine learning models. The key predictors identified are HbA1c, blood glucose, age, and BMI, which are sufficient for accurate diabetes prediction.



	Select	Features	Chi2
1	☑	age	Inf
2	☑	bmi	Inf
3	☑	HbA1c_level	Inf
4	☑	blood_glucose...	Inf
5	☑	hypertension	407.5841
6	☑	smoking_history	344.5559
7	☑	heart_disease	265.2911
8	☑	gender	24.4852

	Select	Features	ANOVA
1	☑	blood_glucose...	Inf
2	☑	HbA1c_level	Inf
3	☑	bmi	Inf
4	☑	age	Inf
5	☑	hypertension	421.8843
6	☑	heart_disease	271.1864
7	☑	smoking_history	173.5618
8	☑	gender	25.9876

	Select	Features	Kruskal Wallis
1	☑	blood_glucose...	Inf
2	☑	HbA1c_level	Inf
3	☑	age	Inf
4	☑	bmi	741.9552
5	☑	hypertension	407.5501
6	☑	heart_disease	265.2691
7	☑	smoking_history	134.9226
8	☑	gender	26.0257

Figure 4 Statistical importance of each predictor using (a) Chi-Square (χ^2), (b) ANOVA, and (c) Kruskal-Wallis algorithms

Table 6 Accuracy of Ensemble Boosted Trees model using different combinations of predictors

Number of predictors	Predictors	Accuracy of validation	Accuracy of the test dataset
1	blood glucose	69.41%	69.27%
1	HbA1c	72.13%	73.27%
1	BMI	65.61%	64.76%
1	Age	42.41%	42.51%
2	HbA1c, blood glucose	82.98%	82.41%
3	HbA1c, blood glucose, BMI	87.96%	88.10%
4	HbA1c, blood glucose, BMI, age	90.79%	90.86%
5	HbA1c, blood glucose, BMI, age, hypertension	91.16%	90.96%
6	HbA1c, blood glucose, BMI, age, hypertension, heart disease	91.06%	91.08%
7	HbA1c, blood glucose, BMI, age, hypertension, heart disease, smoking history	91.14%	91.25%
8	HbA1c, blood glucose, BMI, age, hypertension, heart disease, smoking history, sex	91.49%	91.20%

Table 7 Accuracy of Ensemble Boosted Trees model using only non-blood-based predictors

Number of predictors	Predictors	Accuracy of validation	Accuracy of the test dataset
1	BMI	65.61%	64.76%
1	Age	42.41%	42.51%
1	hypertension	59.63%	59.10%
1	heart disease	56.02%	55.67%
1	smoking history	60.37%	59.33%
1	Sex	52.85%	54.12%
2	BMI, age	74.65%	73.80%
3	BMI, age, and hypertension	75.08%	74.55%
4	BMI, age, hypertension, heart disease	75.30%	74.80%
5	BMI, age, hypertension, heart disease, smoking history	75.73%	74.69%
6	BMI, age, hypertension, heart disease, smoking history, sex	75.92%	75.06%

Table 7 shows the accuracy of the models trained using non-blood parameters, in turn, by adding one clinical feature. The model trained using only two parameters, BMI and age, can achieve an accuracy of 74.65%. Adding one extra feature, hypertension, improves the model by less than 1%.

In contrast, the other two critical predictors are not derived from blood test results but rather from the patient's environmental factors. Therefore, the researcher aims to determine whether it is possible to predict diabetes using only these two environmental predictors without relying on blood test results.

From the results shown in Figure 5, the best-performing model with blood test results was the Ensemble: Boosted Trees model, achieving a test accuracy of 91.45% using eight features. In contrast, the best model without blood test results achieved a test accuracy of 74.65% and 73.80% for the validation and test cases, respectively. The significant difference in accuracy highlights the importance of blood test results as key predictors. Nevertheless, the model without blood test results still achieved reasonable accuracy, suggesting its potential for providing a rough estimate of diabetes risk.

5. Discussion

The results demonstrate that the Ensemble Boosted Trees model achieved the best performance among all machine learning models when applied to the original dataset containing eight predictors. However, this study critically examined models using fewer predictors that demonstrate high potential for diabetes classification without requiring blood test results. These two predictors exhibited importance value of infinity (∞), similar to blood glucose levels and HbA1c, as demonstrated by Chi-Square (χ^2), ANOVA, and the Kruskal-Wallis algorithm. Subsequently, a model using only these two predictors (age and BMI) eliminated the need for blood test results (HbA1c and blood glucose levels). The Ensemble Boosted Trees model outperformed all other models in predicting diabetes without blood test data. All train-test splitting ratios yielded consistent model performance, indicating that the dataset contained sufficient samples for reliable evaluation. This ensures that the model is generalized well to unseen data while maintaining stability in its predictions.

This study highlights and demonstrates the feasibility of non-blood test-based screening, providing a direct comparison to blood test screening using the same dataset. This approach ensures a fair comparison through 24 machine learning models. This study underscores the critical role of blood test results in diabetes prediction, as HbA1c and blood glucose levels achieved a significantly higher accuracy of 91.45%. A non-invasive approach using only age and BMI yielded a reasonable accuracy of 74.65%, making it a valuable tool for preliminary screening. Feature selection proves that increasing the number of predictors by more than four does not significantly enhance model accuracy. However, this research spots age and BMI as the most powerful non-blood predictors. These findings support the feasibility of implementing a simplified, cost-effective model for diabetes risk assessment, particularly in resource-limited settings.

While the non-invasive model demonstrates promise, several limitations warrant consideration. The reduced accuracy compared to blood test-based models (74.65% vs. 91.45%) suggests that non-invasive screening should complement rather than replace traditional diagnostic methods. Future research should investigate the integration of additional non-invasive biomarkers and validate these findings across diverse populations to enhance the model's generalizability and clinical utility.

6. Conclusion

This study evaluated 24 machine learning models for diabetes prediction, comparing blood test-based and non-invasive approaches. The Ensemble Boosted Trees model achieved 91.45% accuracy with eight predictors including blood markers, while a simplified model using only age and BMI achieved 74.65% accuracy. Feature importance analysis revealed that these two non-invasive predictors exhibited statistical significance comparable to blood-based markers.

These findings demonstrate the potential for non-invasive, AI-assisted screening tools as a first-line approach in resource-limited settings. While the accuracy gap (74.65% vs. 91.45%) indicates that non-invasive models should complement rather than replace blood testing, this approach could improve screening accessibility and identify high-risk individuals requiring further evaluation.

Study limitations include the need for external validation across diverse populations and investigation of additional non-invasive parameters to enhance predictive performance. Future research should focus on integrating family history, lifestyle factors, and other readily obtainable metrics to develop more comprehensive non-invasive screening models. This work establishes a foundation for developing accessible, cost-effective diabetes screening tools that could contribute to improved early detection and management globally.

7. Abbreviations

Abbreviation	**Full Term**
AI	Artificial Intelligence
ML	Machine Learning
BMI	Body Mass Index
HbA1c	Hemoglobin A1c
OGTT	Oral Glucose Tolerance Test
χ^2	Chi-Square
ANOVA	Analysis of Variance
PCA	Principal Component Analysis
F1-score	Harmonic mean of precision and recall

8. CRediT Statement

Sasipatcha Hanmanop: Methodology, Investigation, Formal Analysis, Writing – Original Draft, Visualization.
Tatpol Jongsiri: Methodology, Investigation, Formal Analysis, Writing – Original Draft, Visualization.
Kittitat Waiprasit: Writing – Review & Editing.
Suejit Pechprasarn: Conceptualization, Methodology, Investigation, Writing – Review & Editing, Project Administration, Supervision.

9. Acknowledgement

We acknowledge that this research was primarily designed, written, and conducted by the authors, with AI tools used to assist in refining language and presentation. We are grateful for the diabetes dataset obtained from the open-source Kaggle platform; this research would not have been possible without access to this valuable health data resource. We also thank the Rangsit University Research Institute for providing financial support for this study.

10. References

Alam, S., Hasan, M. K., Neaz, S., Hussain, N., Hossain, M. F., & Rahman, T. (2021). Diabetes mellitus: Insights from epidemiology, biochemistry, risk factors, diagnosis, complications, and comprehensive management. *Diabetology*, *2*(2), 36-50. https://doi.org/10.3390/diabetology2020004

Anupongongarch, P., Kaewgun, T., O'Reilly, J. A., & Suraamornkul, S. (2022). Design and construction of a non-invasive blood glucose and heart rate meter by photoplethysmography. *Journal of Current Science and Technology, 12*(1), 89–101. https://doi.org/10.14456/jcst.2022.9

Bergman, M., Abdul-Ghani, M., Neves, J. S., Monteiro, M. P., Medina, J. L., Dorcely, B., & Buysschaert, M. (2020). Pitfalls of HbA1c in the diagnosis of diabetes. *The Journal of Clinical Endocrinology & Metabolism*, *105*(8), 2803-2811. https://doi.org/10.1210/clinem/dgaa372

Boadu, A. A., Yeboah-Manu, M., Osei-Wusu, S., & Yeboah-Manu, D. (2024). Tuberculosis and diabetes mellitus: The complexity of the comorbid interactions. *International Journal of Infectious Diseases*, *146*, Article 107140. https://doi.org/10.1016/j.ijid.2024.107140

Buzzetti, R., Zampetti, S., & Maddaloni, E. (2017). Adult-onset autoimmune diabetes: Current knowledge and implications for management. *Nature Reviews Endocrinology*, *13*(11), 674-686. https://doi.org/10.1038/nrendo.2017.99

Calibo, M. B. T. (2024). Treatment of chronic and severe diabetes mellitus with ketoacidosis in a four-year-old intact female American Pit Bull Terrier. *Asian Journal of Research in Animal and Veterinary Sciences*, *7*(2), 109-121. https://doi.org/10.9734/ajravs/2024/v7i2291

Carmichael, J., Fadavi, H., Ishibashi, F., Shore, A. C., & Tavakoli, M. (2021). Advances in screening, early diagnosis and accurate staging of diabetic neuropathy. *Frontiers in Endocrinology*, *12*, Article 671257. https://doi.org/10.3389/fendo.2021.671257

Chapakiya, I., Traisuwan, A., Chumpong, S., & Chumpong, K. (2025). Follow-up period classification of type 2 diabetes patients using data mining techniques. *Journal of Health Science and Medical Research, 43*(2), Article e20241083. https://doi.org/10.31584/jhsmr.20241083

Dagliati, A., Marini, S., Sacchi, L., Cogni, G., Teliti, M., Tibollo, V., ... & Bellazzi, R. (2018). Machine learning methods to predict diabetes complications. *Journal of Diabetes Science and Technology*, *12*(2), 295-302. https://doi.org/10.1177/1932296817706375

Davidson, K. W., Barry, M. J., Mangione, C. M., Cabana, M., Caughey, A. B., Davis, E. M., ... & US Preventive Services Task Force. (2021). Screening for prediabetes and type 2 diabetes: US preventive services task force recommendation statement. *Jama*, *326*(8), 736-743. https://doi.org/10.1001/jama.2021.12531

Dritsas, E., & Trigka, M. (2022). Data-driven machine-learning methods for diabetes risk prediction. *Sensors*, *22*(14), Article 5304. https://doi.org/10.3390/s22145304

Ghosh, P., Azam, S., Karim, A., Hassan, M., Roy, K., & Jonkman, M. (2021). A comparative study of different machine learning tools in detecting diabetes. *Procedia Computer Science*, *192*, 467-477. https://doi.org/10.1016/j.procs.2021.08.048

Jalilian, H., Javanshir, E., Torkzadeh, L., Fehresti, S., Mir, N., Heidari-Jamebozorgi, M., & Heydari, S. (2023). Prevalence of type 2 diabetes complications and its association with diet knowledge and skills and self-care barriers in Tabriz, Iran: A cross-sectional study. *Health Science Reports*, *6*(2), Article e1096. https://doi.org/10.1002/hsr2.1096

Khanam, J. J., & Foo, S. Y. (2021). A comparison of machine learning algorithms for diabetes prediction. *ICT Express*, *7*(4), 432-439. https://doi.org/10.1016/j.icte.2021.02.004

Liu, H., & Wang, A., Hu, X., Kang, S., Hu, X., Mu, Y., Wang, Y., & Lyu, Z. (2025). The effects of glycated hemoglobin and body mass index on the relationship between the hemoglobin glycation index and the hypoglycemia risk: A moderated mediation analysis. *Metabolism and Target Organ Damage*, *5*, Article 43. https://doi.org/10.20517/mtod.2025.74

Looareesuwan, P., Boonmanunt, S., Thammasudjarit, R., Siriyotha, S., Pattanaprateep, O., Lukkunaprasit, T., Nimitphong, H., Reutrakul, S., Attia, J., McKay, G., & Thakkinstian, A. (2023). Retinopathy prediction in type 2 diabetes: Time-varying Cox proportional hazards and machine learning models. *Informatics in Medicine Unlocked, 40*, Article 101285. https://doi.org/10.1016/j.imu.2023.101285

Mustafa, M. (2023). *A Comprehensive Dataset for Predicting Diabetes with Medical & Demographic Data*. Retrieved from https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset

Pechprasarn, S., Manavibool, L., Supmool, N., Vechpanich, N., & Meepadung, P. (2023). Predicting Parkinson's Disease severity using telemonitoring data and machine learning models: A principal component analysis-based approach for remote healthcare services during the COVID-19 pandemic. *Journal of Current Science and Technology*, *13*(2), 465–485. https://doi.org/10.59796/jcst.V13N2.2023.694

Pechprasarn, S., Srisaranon, N., & Yimluean, P. (2025). Optimizing diabetes prediction: An evaluation of machine learning models through strategic feature selection. *Journal of Current Science and Technology*, *15*(1), Article 75. https://doi.org/10.59796/jcst.V15N1.2025.75

Pippitt, K., Li, M., & Gurgle, H. E. (2016). Diabetes mellitus: Screening and diagnosis. *American Family Physician*, *93*(2), 103-109.

Poorani, K., Balakannan, S. P., & Karuppasamy, M. (2025). Mitigating data imbalance for robust diabetes diagnosis using machine learning and explainable artificial intelligence. *Journal of Current Science and Technology, 15*(3), Article 111. https://doi.org/10.59796/jcst.V15N3.2025.111

Prabhakar, P. K. (2024). Glucose to complications: Understanding secondary effects in diabetes mellitus. *Sumatera Medical Journal*, *7*(2), 87-95. https://doi.org/10.32734/sumej.v7i2.15998

Qin, Y., Wu, J., Xiao, W., Wang, K., Huang, A., Liu, B., ... & Ren, Z. (2022). Machine learning models for data-driven prediction of diabetes by lifestyle type. *International Journal of Environmental Research and Public Health*, *19*(22), Article 15027. https://doi.org/10.3390/ijerph192215027

Rani, K. J. (2020). Diabetes prediction using machine learning. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, *6*(4), 294-305. https://doi.org/10.32628/CSEIT206463

Sarkar, B. K., Akter, R., Das, J., Das, A., Modak, P., Halder, S., ... & Kundu, S. K. (2019). Diabetes mellitus: A comprehensive review. *Journal of Pharmacognosy and Phytochemistry*, *8*(6), 2362-2371.

Sinsophonphap, T., & Thavornsawadi, K. (2022). The cut-off value of HbA1c for prediabetes and diabetes among obese children and adolescents. *Vajira Medical Journal: Journal of Urban Medicine*, *66*(4), 299–310. https://doi.org/10.14456/vmj.2022.30

Skyler, J. S., Bakris, G. L., Bonifacio, E., Darsow, T., Eckel, R. H., Groop, L., ... & Ratner, R. E. (2017). Differentiation of diabetes by pathophysiology, natural history, and prognosis. *Diabetes*, *66*(2), 241-255. https://doi.org/10.2337/db16-0806

Tiwari, D., & Aw, T. C. (2024). The 2024 American Diabetes association guidelines on standards of medical care in diabetes: Key takeaways for laboratory. *Exploration of Endocrine and Metabolic Diseases*, *1*(4), 158-166. https://doi.org/10.37349/eemd.2024.00013

Zhang, J., Zhang, Z., Zhang, K., Ge, X., Sun, R., & Zhai, X. (2023). Early detection of type 2 diabetes risk: Limitations of current diagnostic criteria. *Frontiers in Endocrinology*, *14*, Article 1260623. https://doi.org/10.3389/fendo.2023.1260623