Cite this article: 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

Journal of Current Science and Technology

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

Optimizing Diabetes Prediction: An Evaluation of Machine Learning Models Through Strategic Feature Selection

Suejit Pechprasarn[1,*], Nichapa Srisaranon[2], and Panpatchanan Yimluean[2]

[1]College of Biomedical Engineering, Rangsit University, Pathum Thani 12000, Thailand
[2]Satriwithaya School, Wat Bowon Niwet, Phra Nakhon, Bangkok 10200, Thailand

[*]Corresponding author; E-mail: suejit.p@rsu.ac.th

Received 14 March 2024; Revised 5 April 2024; Accepted 17 April 2024 2024; Published online 23 December 2024

Abstract

Diabetes, a widespread chronic ailment in the United States, imposes significant economic and health burdens, impacting quality of life and life expectancy. This study analyzes a clinical dataset of 253,680 patients from the Behavioral Risk Factor Surveillance System (BRFSS). The dataset encompasses 21 predictors, including high blood pressure, cholesterol, body mass index (BMI), smoking, stroke, heart disease, physical activity, fruit consumption, vegetable consumption, alcohol consumption, insurance coverage, lack of medical visits due to financial constraints, general health, days with mental health issues, days with physical injuries in the past 30 days, difficulties in walking, gender, age, income, and education level. The objective is to balance the training dataset, compare different supervised machine learning models, and identify critical clinical features contributing to diabetes using unsupervised feature selection methods. A total of 34 machine learning models in MATLAB2023a were trained and compared. Quadratic Support Vector Machine (SVM), Coarse Gaussian SVM, and Narrow Neural Networks achieved the highest training accuracy (76.3%), while the Bilayered Neural Network attained 74.7% on an unseen test dataset. Among all, Quadratic SVM demonstrated the best overall performance based on average accuracy, precision, recall, and F1 score. Feature selection highlighted nine key predictors: high blood pressure, high cholesterol, BMI, heart disease, physical activity, general health, recent bodily injuries, mobility issues, and age. A model trained on these features achieved a commendable accuracy of 75.4%, demonstrating the feasibility of a simplified, efficient diagnostic tool with a diagnostic efficacy of 0.7.

This study underscores the potential of streamlined models to predict diabetes with fewer parameters while maintaining high accuracy, offering a valuable tool for healthcare diagnostics.

Keywords: *diabetes diagnosis; machine learning; feature selection; complexity-reduced model; intelligent diagnostic software*

1. Introduction

Diabetes mellitus, a metabolic condition marked by high blood glucose levels, is significant in medical history, with roots tracing back to ancient civilizations. The word "diabetes" derives from a Greek term meaning "siphon," which refers to the observation of sweet-tasting urine in people with the disease (Anupongongarch et al., 2022). Diabetic patients were notably recorded by Apollonius of Memphis around 250 to 300 BC, who documented this peculiar symptom, marking one of the earliest accounts of diabetes (Ahmed, 2002).

Throughout the centuries, our understanding of diabetes has evolved, propelled by critical discoveries and advancements. One pivotal moment occurred in 1922 when Banting, Best, and Collip successfully isolated insulin from cow pancreas, introducing a life-saving therapy for individuals with diabetes. Despite such breakthroughs, diabetes remains a formidable global health challenge, ranking among the top causes of morbidity and mortality worldwide (Sims et al., 2021).

The landscape of diabetes encompasses various forms, each with its unique etiology, clinical presentation, and management considerations. Type 1 diabetes is defined by the autoimmune obliteration of cells that produce insulin, typically manifesting in youth but can occur at any age. On the other hand, Type 2 diabetes, frequently linked with insulin resistance and lifestyle factors, primarily impacts adults but is progressively identified in children and teenagers (Niramitmahapanya et al., 2023). Gestational diabetes, arising during pregnancy, poses risks to maternal and fetal health, underscoring the importance of timely intervention (Popoviciu et al., 2023).

Emerging subtypes, such as Type 3c and maturity-onset diabetes of the young (MODY), further diversify the diabetes spectrum, highlighting the complexity of this condition. Type 3c diabetes results from pancreatic damage unrelated to autoimmunity, while MODY stems from genetic mutations affecting insulin production. Acknowledging these nuances is essential for crafting individualized treatment approaches and ultimately enhancing patient outcomes (Hart et al., 2016).

Despite medical advancements, the prevalence of diabetes continues to escalate globally, driven by factors like sedentary lifestyles, poor dietary habits, and rising obesity rates. In the United States, approximately 37.3 million people are living with diabetes, with the bulk of these cases being Type 2 diabetes. Globally, an estimated 537 million adults grapple with diabetes, with projections indicating a steady rise in prevalence in the coming decades (Sugandh et al., 2023).

The pathophysiology of diabetes revolves around the dysregulation of glucose metabolism, leading to hyperglycemia and subsequent complications that have an impact on multiple organ systems. Insulin therapy remains a cornerstone in Type 1 diabetes management, while lifestyle modifications are pivotal in Type 2 diabetes prevention and treatment (Manosroi et al., 2023; Banday et al., 2020).

Recent research has delved into innovative treatments, including verapamil therapy and the use of artificial intelligence, to enhance diabetes management. These advancements promise to improve patient outcomes and address the evolving challenges posed by diabetes (Guan et al., 2023).

Applying Machine Learning (ML) techniques for predicting diabetes has gathered significant attention in recent research, utilizing various algorithms and datasets to enhance prediction accuracy and early detection capabilities. Recent studies focus on the importance of selecting appropriate ML algorithms, preprocessing techniques, and feature engineering to address the challenges of diabetes prediction effectively (Panda et al., 2024). Recent studies have compared the efficiency of multiple ML algorithms like Random Forest, K-Nearest Neighbors (KNN), and multilayer perceptrons in predicting diabetes, showing the potential of these algorithms to achieve high accuracy rates. For instance, Almahdawi et al., (2022) demonstrated that Random Forest has superior efficiency classifiers with a 98.8% accuracy rate using a dataset of Iraqi patients (Almahdawi et al., 2022).

A critical comparison of ML techniques reveals the superiority of specific algorithms in certain contexts. For instance, a study comparing the performance of various ML algorithms found that Logistic Regression and Support Vector Machines showed promising results in diabetes prediction, emphasizing the importance of algorithm selection based on the dataset characteristics (Khanam, & Foo, 2021).

The application of ML for the early detection of diabetes is focused as a key advantage, allowing for timely intervention and management. Lu et al., (2023) achieved an accuracy of 99.1% using the XGBoost classifier, demonstrating the potential for early and accurate diabetes prediction (Lu et al., 2023).

The exploration of ensemble methods and novel ML approaches for diabetes prediction shows promising results, with studies reporting improved accuracy and predictive performance. For example, the fusion of Support Vector Machine and Artificial Neural Network models demonstrated a prediction accuracy of 94.87%, underscoring the effectiveness of combined techniques in disease prediction (Ahmed et al., 2022).

Therefore, recent advancements in ML for diabetes prediction are marked by the exploration of various algorithms, the critical role of data preprocessing, and the adoption of innovative techniques such as ensemble methods. These studies underscore the potential of ML to revolutionize diabetes prediction, facilitating early detection and contributing to more effective disease management strategies.

After examining numerous studies, it's evident that the accuracy of forecasting tools for type 2 diabetes is remarkably impressive. The magic number, the area under the curve (AUC), ranged from .7182 to .7949 across the board. Now, if you are talking sheer precision, the neural network model stole the show with an accuracy of 82.4%, specificity hitting 90.2%, and an AUC that topped the charts at .7949. But here is a twist: when pinpointing type 2 diabetes with the finesse of a detective, the decision tree model was the one to beat, boasting the highest sensitivity at 51.6%. The plot thickens with lifestyle and health habits playing a pivotal role. Individuals clocking in 9 or more hours of sleep nightly found themselves at a slightly higher risk, with an adjusted odds ratio (aOR) of 1.13 and a confidence interval that whispered a tale of caution (1.03–1.25). And those who made less frequent visits to their doctor than once a year? Their risk skyrocketed, with an aOR of 2.31 and a 95% confidence interval screaming warning signs (1.86–2.85). Among the contenders, eight models were scrutinized, and though the neural network emerged as the champion in AUC regarding the front lines of type 2 diabetes screenings, the decision tree model takes the cake. Its unmatched sensitivity offers hope for catching the condition early on (Xie ZiDian et al., 2019).

Despite the considerable progress in applying machine learning within the healthcare sector, particularly in diagnosing chronic diseases like diabetes, several gaps remain in the current research landscape. Firstly, there is a noticeable deficiency in studies that systematically compare the performance of various ML models using the same dataset, which includes a wide range of clinical variables and patient characteristics. The absence of comprehensive comparisons makes it challenging to determine the most effective models for diabetes diagnosis.

Secondly, existing research has insufficient consideration for the interpretability of ML models in clinical settings. Understanding and explaining predictions is crucial for clinical acceptance and application, yet many studies focus primarily on predictive accuracy without sufficient emphasis on how these predictions are derived and how they can be integrated into clinical workflows.

Furthermore, while numerous studies have applied ML models to diabetes diagnosis, there is a significant variation in the quality and diversity of datasets used, leading to potentially biased or non-generalizable findings.

Another critical gap is the limited exploration of the impact of patient characteristics on the accuracy of diabetes diagnoses. Understanding how different factors influence model predictions can provide insights into disease mechanisms and help tailor diagnostic processes to individual patients, enhancing personalized medicine.

Finally, there is a scarcity of research that benchmarks newly developed ML models against existing diagnostic standards and models reported in the literature. Such benchmarking is essential to demonstrate the added value of new models and justify their implementation in clinical practice.

Our study aims to fill these gaps by conducting a rigorous comparative analysis of statistical machine learning models. We emphasize accuracy and prioritize model interpretability to facilitate seamless integration into clinical practice. By using a comprehensive and diverse dataset for training and employing a 5-fold cross-validation approach, we intend to ensure the robustness and generalizability of our findings. Furthermore, our research will explore the impact of patient characteristics on model accuracy, offering valuable insights for personalized medicine approaches. By setting clear benchmarks for model performance comparison with existing literature, our study seeks to make a significant contribution to the field of diabetes diagnosis using ML, paving the way for more effective and personalized diagnostic processes.

The behavioral risk factor surveillance system (BRFSS), a vital tool in public health surveillance, offers crucial data on diabetes prevalence and associated risk factors. All 21 predictors were utilized in this study. The objective is to reduce this number to the minimum number of parameters that can still yield substantial accuracy. Consequently, the parameter count is reduced to 9, although this reduction entails a trade-off in a slight decrease in classification performance.

2. Objectives

The main objective of this study is to employ statistical machine learning models to identify key clinical variables crucial for diagnosing diabetes and evaluate the accuracy of these models in categorizing diabetes cases, while comparing the predicted accuracy rates with those reported in existing literature to provide insights into the effectiveness of ML models in the context of diabetes.

3. Materials and Methods

This section elaborates on the methodology, including the software utilized, data source, data management procedures, training dataset and testing dataset preparation, supervised classification training, and feature selection tool. The process flow of this research is depicted in Figure 1 and provided in the subsections below.

3.1 Dataset

The Diabetes Health Indicators dataset analyzed here was obtained from the open-source Kaggle (Teboul, 2022) (assessed January 2, 2024). It contains data from 253,680 survey responses to identify diabetes, including the following 22 attributes, which comprise 21 predictors and 1 label. These features include questions directly asked of participants or calculated variables based on individual participant responses. The 22 attributes are shown in Table 1 below. The 21 predictors have 2 classes, 0 indicating a non-diabetic person, whereas 1 indicates prediabetes and diabetes cases. This dataset has 21 clinical predictors and is not balanced. While the dataset utilized in our study is available under a Creative Commons (CC) license, which permits the use and analysis of the data, we recognize that ethical considerations in research extend beyond legal licensing agreements. To this end, even though the CC license provides certain freedoms concerning the use of the data, we proactively sought to adhere to the highest ethical standards related to the protection of patient data. This included removing all potentially identifiable information from the dataset before analysis and ensuring that our research practices strictly complied with ethical guidelines for human subject's research, affirming our commitment to ethical research practices and protecting individual privacy.

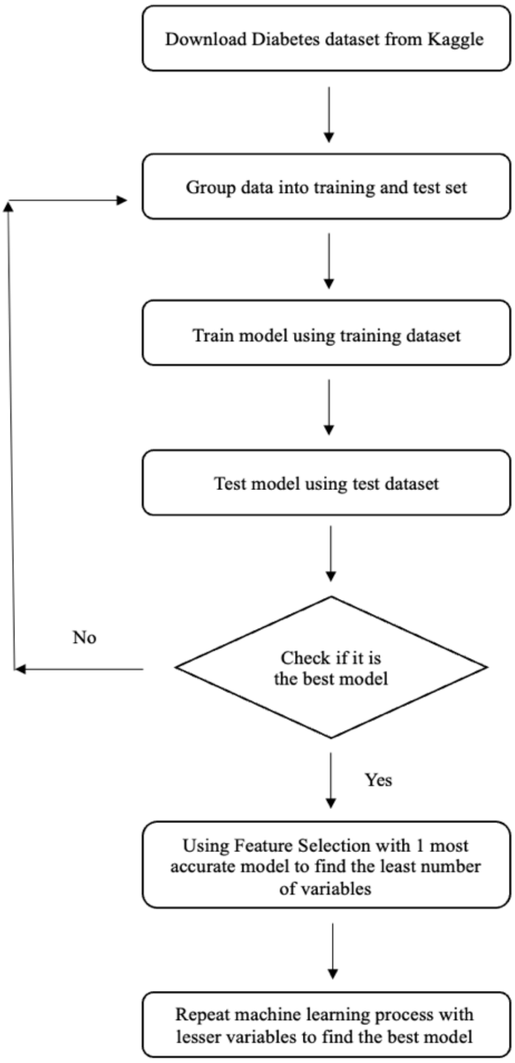
Download Diabetes dataset from Kaggle
Group data into training and test set
Train model using training dataset
Test model using test dataset
Check if it is the best model
No
Yes
Using Feature Selection with 1 most accurate model to find the least number of variables
Repeat machine learning process with lesser variables to find the best model

Figure 1 Process flow of the ML analysis in this research

3.2 Data Curation

In this study, we utilized a dataset comprising 253,680 rows without any reduction despite inherent bias. The dataset was divided into a training dataset and a testing dataset. The training dataset consists of 67,158 rows. Within this training dataset, 33,579 rows were dedicated to cases with no diabetes (0), while the remaining 33,579 rows represented prediabetes or diabetes cases (1). Additionally, a test dataset was formed, containing 186,522 rows. The test dataset has 184755 rows of no diabetes cases (0). However, the other 1767 rows represented prediabetes or diabetes cases (1). Notably, this test dataset had a bias towards the no diabetes (0) case, which is unsuitable for machine learning training since bias testing may cause the machine learning to predict more output of no diabetes than prediabetes or diabetes cases. However, we opted not to decrease the dataset size, ensuring training was conducted using a balanced dataset, while acknowledging that the test dataset could be imbalanced.

3.3 Dataset for Training and Testing

The 253,680 rows were then divided into 2 datasets: the training dataset (67,158 rows) and the test dataset (186,522 rows) at a ratio of 95% to 5%. The training and test datasets were selected randomly by ensuring that training datasets contained the same number of 0 case and 1 case, in other words, unbiased training datasets and biased test datasets.

Table 1 Predictors and labels obtained from the Kaggle Diabetes Health Indicators dataset

Predictors	**Details**
High BP	0: no high blood pressure, 1: high blood pressure
HighChol	0: no high cholesterol, 1: high cholesterol
CholCheck	0: no cholesterol check in 5 years, 1: yes cholesterol check in 5 years
BMI	Body mass index
Smoker	Have you ever smoked more than 100 cigarettes in your life? 0: No, and 1: Yes
Stroke	(Ever told) you had a stroke. 0: No, and 1: yes
HeartDiseaseorAttack	Diagnosed for myocardial infarction (MI) or coronary heart disease (CHD)? 0: No, and 1: Yes
PhysActivity	Physical activity in the past 30 days, excluding physical activities due to occupation. 0: No, and 1: Ye
Fruits	Consume some fruits at least one time per day. 0: No, and 1: Yes.
Veggies	Consume some vegetables at least one time per day. 0: No, and 1: Yes.
HvyAlcoholConsump	Adult men consumed at least 14 drinks per week, and adult women consumed at least 7 drinks per week. 0: No, and 1: Yes.
AnyHealthcare	Have any health care coverage, including health insurance and prepaid plans? 0: No, and 1: Yes.
NoDocbcCost	Was there a time in the past 12 months when you needed to see a doctor but could not because of financial constraints? 0: No, and 1: Yes.
GenHlth	Would you say that, in general, your health is on a scale of 1-5? 1 means excellent, 2 means very good, 3 means good, 4 means fair, and 5 means poor.
MentHlth	Days of poor mental health: 1 day to 30 days.
PhysHlth	Physical illness or injury days in the past 30 days on a scale of 1-30.
DiffWalk	Do you experience significant challenges with walking or ascending stairs? 0: No, and 1: Yes
Sex	0: female, and 1: male
Age	13-level age category (AGEG5YR see codebook) 1 = 18-24 9 = 60-64 13 = 80 or older
Education	Education level (EDUCA see codebook) scale 1-6 1 = Never attended school or only kindergarten 2 = elementary.
Income	Income scale (INCOME2 see codebook) scale 1-8 1 = less than $10,000 5 = less than $35,000 8 = $75,000 or more
Label	**Details**
Diabetes_binary	0 = no diabetes 1 = prediabetes or diabetes

3.4 Machine Learning Training

The built-in Classification Learner in MATLAB 2022b was utilized to train 34 machine learning models using a training dataset. These methods encompass Fine Tree, Medium Tree, Coarse Tree, Linear Discriminant Analysis, Quadratic Discriminant Analysis, Binary Generalized Linear Model (GLM) Logistic Regression, Efficient Logistic Regression, Efficient Linear Support Vector Machine (SVM), Gaussian Naive Bayes, Kernel Naive Bayes, Linear SVM, Quadratic SVM, Cubic SVM, Fine Gaussian SVM, Medium Gaussian SVM, Coarse Gaussian SVM, Fine K-Nearest Neighbors (KNN), Medium KNN, Coarse KNN, Cosine KNN, Cubic KNN, Weighted KNN, Boosted Trees, Bagged Trees, Subspace Discriminant Analysis, Subspace KNN, RUSBoosted Trees, Narrow Neural Network, Medium Neural Network, Wide Neural Network, Bilayer Neural Network, Trilayer Neural Network, Kernel SVM, and Logistic Regression Kernel. Numerous models were trained to compare their performance for the given dataset. After training, the models were tested using a test dataset to determine whether the trained models were generalized and could provide correct response for the unseen dataset. Since the test dataset is biased, the models' performance was assessed by computing classification accuracy. All the available supervised machines in MATLAB 2022b were employed to choose the best model out of 34 machine learning models because there is no one-size-fits-all solution (Lyngdoh et al., 2021). Depending on the dataset, the best-performing model may vary. In this case, the best model for predicting diabetes performs well.

The performance of the classification task of the listed models was evaluated using 5-fold cross-validation, measuring accuracy, precision, recall, and F1 score, as detailed in Equations (1) through (4).

$$\text{Precision}=\frac{T_p}{T_p+F_p} \tag{1}$$

$$\text{Recall}=\frac{T_p}{T_p+F_n} \tag{2}$$

$$F_1=2\,\frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \tag{3}$$

$$\text{Accuracy}=\frac{T_p+T_n}{T_p+T_n+F_p+F_n} \tag{4}$$

where T_p and T_n represent true positive and true negative cases, respectively.

F_p and F_n represent false positive and false negative cases, respectively.

3.5 Feature Selection

Feature Selection was applied to ascertain the optimal number of crucial variables from the set of 21 predictors. Feature selection is an analysis to identify which predictors significantly contribute to the output function. Feature ranking algorithm, including MRMR, Chi^2, reliefF, ANOVA, and Kruskal-Wallis, were applied for the unsupervised ranking task based on the importance of the 21 predictors from highest to lowest accuracy. The predictors were then trained from highest to lowest accuracy according to each feature ranking algorithm to determine the minimum number of predictors required, which maintains the accuracy with the closest when using 21 predictors.

4. Results

4.1 Models Training using all 21 Predictors and Training Dataset

The 34 ML models were first trained using all 21 predictors with no diabetes (0) and prediabetes or diabetes (1) labels, as listed in Table 1. The accuracy of each model is shown in Table 2. The top three ML models were the Quadratic SVM (SVM), Coarse Gaussian SVM (SVM), and Narrow Neural Network (Neural Network), with the same accuracy of 76.3%. The confusion matrices corresponding to the top three ML models are illustrated in Figures 2a, 2b, and 2c, respectively. These data are utilized to compute validation accuracy, precision, recall, and the F1 score by employing Equations (1) through (4) and referencing the confusion matrices depicted in Figure 2.

Table 2 Classification validation accuracy of models trained with all 21 predictors using a 5-fold cross-validation method

Model	Details	Accuracy	Precision	Recall	F$_1$ Score	Average
Tree	Fine Tree	74.9%	72.6%	79.9%	76.1%	75.9%
	Medium Tree	73.8%	72.3%	77.1%	74.6%	74.5%
	Coarse Tree	71.2%	69.2%	76.5%	72.6%	72.4%
Linear Discriminant	Linear Discriminant	75.8%	74.4%	78.5%	76.4%	76.3%
Quadratic Discriminant	Quadratic Discriminant	73.5%	71.5%	78.4%	74.8%	74.5%
Binary GLM Logistic Regression	Binary GLM Logistic Regression	75.9%	75.0%	77.8%	76.4%	76.3%
Efficient Logistic Regression	Efficient Logistic Regression	75.9%	75.0%	77.8%	76.4%	76.3%
Efficient Linear SVM	Efficient Linear SVM	75.9%	73.9%	80.0%	76.9%	76.7%
Naive Bayes	Gaussian Naive Bayes	72.6%	73.1%	71.5%	72.3%	72.4%
	Kernel Naïve Bayes	73.5%	68.9%	85.4%	76.3%	76.0%
SVM	Linear SVM	75.9%	73.9%	79.9%	76.8%	76.6%
	Quadratic SVM	76.3%	72.8%	83.8%	77.9%	77.7%
	Cubic SVM	75.6%	73.7%	79.7%	76.6%	76.4%
	Fine Gaussian SVM	70.3%	65.8%	84.4%	74.0%	73.6%
	Medium Gaussian SVM	76.2%	73.7%	81.3%	77.3%	77.1%
	Coarse Gaussian SVM	76.3%	73.4%	82.6%	77.7%	77.5%
KNN	Fine KNN	68.0%	68.1%	67.7%	67.9%	67.9%
	Medium KNN	73.2%	73.3%	73.0%	73.1%	73.2%
	Coarse KNN	75.1%	73.1%	79.6%	76.2%	76.0%
	Cosine KNN	73.3%	73.5%	72.9%	73.2%	73.2%
	Cubic KNN	73.2%	73.2%	73.0%	73.1%	73.1%
	Weighted KNN	73.0%	71.4%	76.8%	74.0%	73.8%
Ensemble	Boosted Trees	75.8%	73.6%	80.4%	76.9%	76.7%
	Bagged Trees	73.4%	72.5%	75.2%	73.8%	73.7%
	Subspace Discriminant	75.3%	74.6%	76.8%	75.7%	75.6%
	Subspace KNN	56.2%	85.2%	15.0%	25.5%	45.5%
	RUSBoosted Trees	73.8%	72.3%	77.1%	74.6%	74.5%
Neural Network	Narrow Neural Network	76.3%	74.2%	80.6%	77.3%	77.1%
	Medium Neural Network	76.0%	73.9%	80.3%	77.0%	76.8%
	Wide Neural Network	74.6%	73.3%	77.4%	75.3%	75.1%
	Bilayered Neural Network	76.1%	74.0%	80.4%	77.1%	76.9%
	Trilayered Neural Network	76.0%	74.1%	80.2%	77.0%	76.8%
Kernel	SVM Kernel	73.2%	70.8%	79.1%	74.7%	74.4%

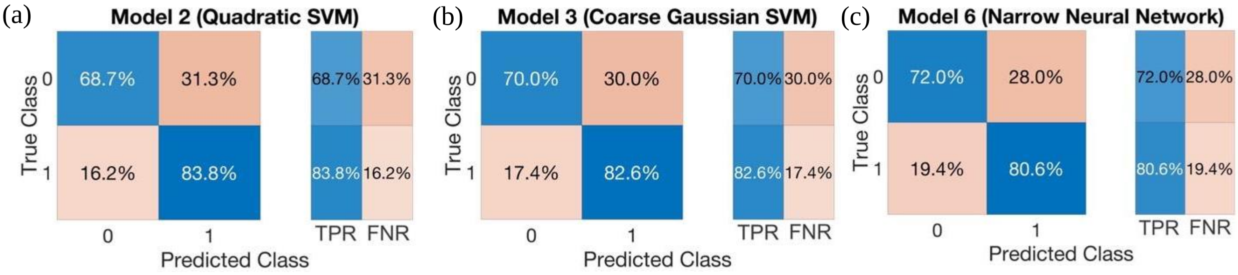

Figure 2 Confusion matrices of top three trained models (a) Quadratic SVM, (b) Coarse Gaussian SVM, and (c) Narrow Neural Network

The statistical analysis was expanded to incorporate confidence intervals and effect sizes for the performance metrics of the assessed machine learning models to offer a more detailed understanding of the findings. For each model, 95% confidence intervals were computed for accuracy, precision, recall, and F1 scores using bootstrap methods with 1,000 samples, ensuring an accurate representation of underlying metric distributions. This analysis not only enhances the reliability of our results but also aids in discerning the practical significance of the performance differences observed among the models. Including these statistical measures helps clarify the trustworthiness and efficacy of the models in predicting diabetes outcomes, ensuring a comprehensive evaluation suited for clinical decision-making.

At 95% confidence, the Quadratic SVM demonstrated performance metric ranges of [78.9%, 82.3%] for accuracy, [79.9%, 84.4%] for precision, [79.9%, 84.4%] for recall, and [80.3%, 83.8%] for the F1-score. The Coarse Gaussian SVM showed performance ranges of [81.0%, 84.2%] for accuracy, [82.6%, 86.6%] for precision, [82.2%, 86.6%] for recall, and [83.0%, 86.1%] for the F1-score. Similarly, the Narrow Neural Network achieved impressive ranges of [82.1%, 85.4%] for accuracy, [83.8%, 87.9%] for precision, [83.8%, 87.9%] for recall, and [84.3%, 87.5%] for the F1-score at 95% confidence intervals.

4.2 Models Testing using all 21 Predictors and the Separated Test Dataset

Table 3 Classification Accuracy of ML Models Tested with all 21 Predictors

Model	Details	Accuracy	Precision	Recall	F_1 Score	Average
Tree	Fine Tree	73.3%	70.4%	80.1%	75.0%	74.7%
	Medium Tree	72.6%	71.7%	74.6%	73.1%	73.0%
	Coarse Tree	70.0%	67.9%	75.7%	71.6%	74.2%
Linear Discriminant	Linear Discriminant	73.6%	72.2%	76.8%	74.4%	74.2%
Quadratic Discriminant	Quadratic Discriminant	71.8%	69.6%	77.4%	73.3%	73.0%
Binary GLM Logistic Regression	Binary GLM Logistic Regression	73.9%	72.7%	76.5%	74.6%	74.4%
Efficient Logistic Regression	Efficient Logistic Regression	73.9%	72.7%	76.5%	74.6%	74.4%
Efficient Linear SVM	Efficient Linear SVM	73.9%	71.7%	78.7%	75.1%	74.8%
Naive Bayes	Gaussian Naive Bayes	70.9%	71.3%	70.0%	70.6%	70.7%
	Kernel Naïve Bayes	71.5%	67.0%	84.6%	74.8%	74.5%
SVM	Linear SVM	73.9%	71.7%	78.7%	75.1%	74.8%
	Quadratic SVM	74.3%	70.6%	83.1%	76.3%	76.1%*
	Cubic SVM	74.0%	71.8%	78.8%	75.2%	74.9%
	Fine Gaussian SVM	68.5%	64.4%	82.3%	72.3%	71.9%
	Medium Gaussian SVM	74.3%	71.5%	80.9%	75.9%	75.6%
	Coarse Gaussian SVM	74.5%*	71.2%	82.2%	76.3%	76.0%
KNN	Fine KNN	66.0%	65.7%	66.8%	66.3%	66.2%
	Medium KNN	71.3%	71.2%	71.3%	71.3%	71.3%
	Coarse KNN	73.1%	70.9%	78.2%	74.4%	74.1%
	Cosine KNN	71.2%	71.2%	71.1%	71.1%	71.1%
	Cubic KNN	71.0%	70.9%	71.0%	71.0%	71.0%
	Weighted KNN	70.4%	68.8%	74.6%	71.6%	71.4%
Ensemble	Boosted Trees	74.0%	71.5%	79.9%	75.4%	75.2%
	Bagged Trees	71.9%	70.5%	75.1%	72.7%	72.6%
	Subspace Discriminant	73.3%	72.5%	75.0%	73.7%	73.6%
	Subspace KNN	54.7%	84.6%	11.5%	0.2%	37.8%
	RUSBoosted Trees	72.6%	71.7%	74.6%	73.1%	73.0%
Neural Network	Narrow Neural Network	74.4%*	71.9%	80.0%	75.7%	75.5%
	Medium Neural Network	74.2%	71.8%	79.6%	75.5%	75.3%
	Wide Neural Network	72.6%	70.9%	76.7%	73.7%	73.5%
	Bilayered Neural Network	74.7%*	72.1%	80.5%	76.1%	75.9%
	Trilayered Neural Network	74.4%*	71.9%	80.0%	75.7%	75.5%
Kernel	SVM Kernel	72.0%	69.4%	78.7%	73.8%	73.5%
	Logistic Regression Kernel	70.8%	68.8%	75.9%	72.2%	71.9%

The separated test dataset was then employed to predict the classification output compared to its known label. The classification accuracy of the evaluated 34 models using 21 predictors and the test dataset is shown in Table 3. The top 4 models that obtained the highest accuracy were Bilayered Neural Network (Neural Network) with an accuracy of 74.7%, Coarse Gaussian SVM (SVM) with an accuracy of 74.5%, Narrow Neural Network (Neural Network) and Trilayered Neural Network (Neural Network) with an accuracy of 74.4%. The Quadratic SVM (SVM), Coarse Gaussian SVM (SVM), and Narrow Neural Network (Neural Network) that achieved the highest accuracy in the trained model of 76.3% were now reduced their accuracy to 74.3%, 74.5%, and 74.4%, respectively in the testing. However, the model that obtained the highest average of classification performance was Quadratic SVM (SVM) with 76.1%; therefore, Quadratic SVM (SVM) was the most reliable model. The confusion matrix of the model, when tested using the unseen test dataset, is shown in Figure 3. The tested model of Quadratic SVM exhibited performance metric ranges of [81.6%, 84.8%] for accuracy, [83.7%, 87.8%] for precision, [83.7%, 87.6%] for recall, and [84.3%, 87.2%] for the F1-score at 95% confidence intervals. The range of the test performance metrics at a 95% confidence interval was similar to the training performance, with a slight underfitting within 3%.

4.3 Number of Predictors Reduction using Feature Selection

In the previous sections, we have demonstrated that accurate classification models can be trained using all the available 21 predictors, which are (1) HighBP, (2) HighChol, (3) Cholcheck, (4) BMI, (5) Smoker, (6) Stroke, (7) HeartDiseaseorAttack, (8) PhysActivities, (9) Fruits, (10) Veggies, (11) HvyAlcoholConsump, (12) AnyHealthcare, (13) NoDocbsCost, (14) GenHLth, (15) MentHlth, (16) PhysHlth, (17) DiffWalk, (18) Sex, (19) Age, (20) Education, (21) Income. Here, we performed principal component analysis (PCA), but since the dataset had numerous data rows, an error occurred; therefore, feature selection analysis in the built-in Matlab feature selection was used to identify crucial predictors based on unsupervised statistical approaches.

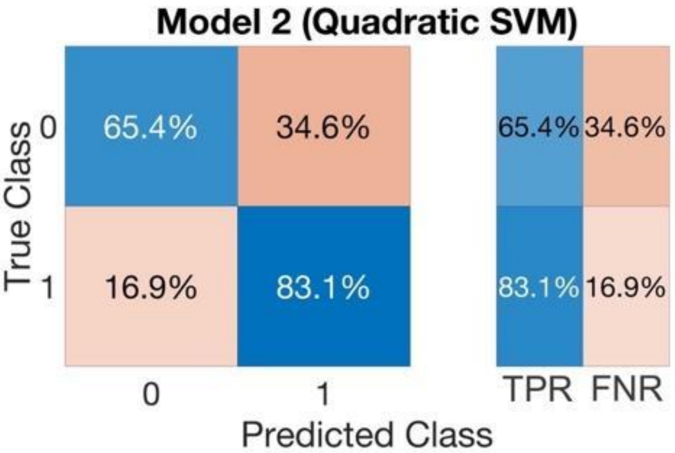

Figure 3 Confusion matrix of tested Quadratic SVM model

Table 4 statistical values of each predictor from each algorithm within the feature selection tool

No.	MRMR		Chi^2		ReliefF		ANOVA		Kruskal Wallis	
1	GenHlth	0.0964	HighBP	Inf	GenHlth	0.0098	Income	Inf	Income	Inf
2	CholCheck	0.079	HighChol	Inf	Age	0.008	Education	Inf	Education	Inf
3	Age	0.0584	BMI	Inf	Income	0.0057	Age	Inf	Age	Inf
4	HvyAlcoholConsump	0.0458	HeartDiseaseorAttack	Inf	BMI	0.0056	DiffWalk	Inf	DiffWalk	Inf
5	BMI	0.0285	PhysActivity	Inf	PhysHlth	0.0047	PhysHlth	Inf	PhysHlth	Inf
6	HighChol	0.0228	GenHlth	Inf	Education	0.0033	GenHlth	Inf	GenHlth	Inf

Table 4 Cont.

No.	MRMR		Chi^2		ReliefF		ANOVA		Kruskal Wallis	
7	Veggies	0.0216	PhysHlth	Inf	HvyAlcoholConsump	0.0024	PhysActivity	Inf	PhysActivity	Inf
8	HighBP	0.0214	DiffWalk	Inf	CholCheck	0.0022	HeartDiseaseorAttack	Inf	HeartDiseaseorAttack	Inf
9	Stroke	0.0182	Age	Inf	Stroke	0.0021	BMI	Inf	BMI	Inf
10	Income	0.0175	Education	Inf	HighBP	0.0015	HighChol	Inf	HighChol	Inf
11	HeartDiseaseorAttack	0.0173	Income	Inf	HeartDiseaseorAttack	0.0013	HighBP	Inf	HighBP	Inf
12	PhysActivity	0.015	Stroke	527.184	HighChol	0.0012	Stroke	531.2916	Stroke	527.1762
13	DiffWalk	0.0125	CholCheck	412.902	Sex	0.0009	CholCheck	415.4049	CholCheck	412.8959
14	Smoker	0.0084	HvyAlcoholConsump	345.8317	NoDocbcCost	0.0009	HvyAlcoholConsump	347.5785	HvyAlcoholConsump	345.8266
15	Education	0.0082	Veggies	344.2773	DiffWalk	0.0008	Veggies	346.0081	Veggies	344.2722
16	PhysHlth	0.0072	MentHlth	306.9365	PhysActivity	0.0005	Smoker	261.9452	Smoker	260.9559
17	Fruits	0.0056	Smoker	260.9597	Fruits	0.0004	MentHlth	235.0599	Fruits	182.3295
18	MentHlth	0.0028	Fruits	182.3322	Smoker	0.0001	Fruits	182.8065	Sex	56.2867
19	Sex	0.0021	Sex	56.2875	MentHlth	0	Sex	56.3289	MentHlth	51.6536
20	AnyHealthcare	0.0017	NoDocbcCost	40.9089	AnyHealthcare	-0.0001	NoDocbcCost	40.9299	NoDocbcCost	40.9084
21	NoDocbcCost	0.001	AnyHealthcare	27.8667	Veggies	-0.0005	AnyHealthcare	27.8758	AnyHealthcare	27.8663

Table 5 Accuracy of Quadratic SVM model using a different order of predictors from different algorithm

Number of Predictors	MRMR	Chi^2	ReliefF	ANOVA	Kruskal-Wallis
1	47.5%	69.6%	47.5%	52.2%	52.2%
2	59.5%	69.6%	63.5%	47.1%	47.1%
3	50.2%	52.0%	53.0%	51.6%	51.6%
4	48.3%	58.2%	55.5%	55.7%	55.7%
5	48.9%	59.9%	60.4%	60.5%	60.5%
6	55.3%	69.0%	68.5%	67.9%	67.9%
7	60.7%	72.4%	71.3%	70.6%	70.6%
8	71.2%	71.6%	70.8%	70.8%	70.8%
9	72.3%	75.4%	74.3%	74.7%	74.7%
10	75.5%	75.5%	75.3%	75.2%	75.2%
11	75.6%	75.7%	75.5%	75.7%	75.7%

The feature selection tool utilized five algorithms: MRMR, Chi^2, ReliefF, ANOVA, and Kruskal Wallis. The statistical values of each predictor from each algorithm are shown below in Table 4. The 21 predictors were ranked in descending order to indicate which predictor is the most important among all the predictors.

In Table 4, each algorithm shares common top-ranking predictors, with some having infinite values. All algorithms were utilized to train quadratic SVM models to determine the best algorithm, using a range of predictors from one to eleven, following the order specified by each algorithm. This research process is time-consuming, requiring 2 hours to train a single model, and 55 models were trained as part of this study. The accuracy of Quadratic SVM models is shown in Table 5.

According to Table 5, among the five ranking algorithms, including MRMR, Chi^2, ReliefF, ANOVA, and Kruskal Wallis, the result of using Chi^2 with 9 predictors demonstrates the highest accuracy of 75.4% followed by ANOVA, and Kruskal Wallis with an accuracy of 74.7%. Here, the 9 predictors were chosen because Chi^2 gave the highest accuracy among the other ranking algorithms. The confusion matrix for the Quadratic SVM model trained using 9 predictors is shown below in Figure 4.

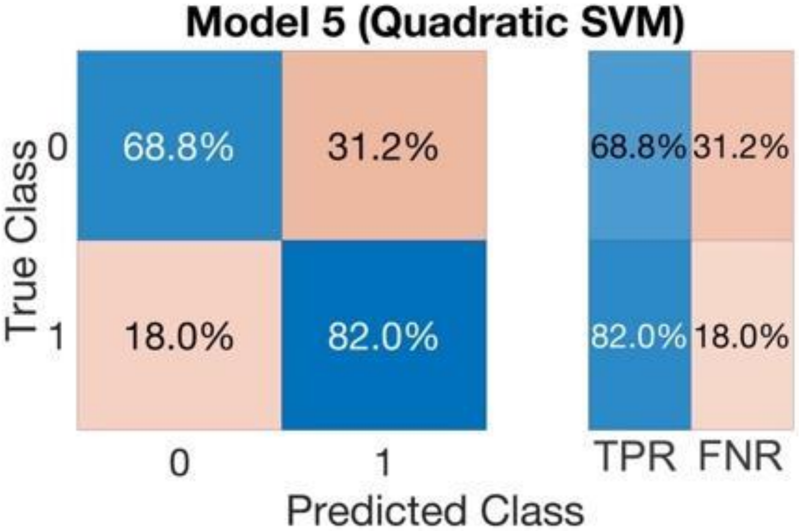

Figure 4 The confusion matrix of the Quadratic SVM model using 9 predictors

The 9 highest accuracy predictors, namely HighBP, HighChol, BMI, HeartDisease, PhysActivity, GenHlth, PhysHlth, DiffWalk, and Age, were through the feature selection process. Based on each ranking algorithm, the predictors were then trained in sequence from highest to lowest accuracy. This approach aimed to identify the minimum number of predictors that could maintain accuracy closest to that achieved using all 21 predictors. The outcome revealed that employing only nine predictors from the Chi^2 ranking algorithm resulted in a 75.4% accuracy, closely resembling the percentage achieved with the complete set of 21 predictors. At 95% confidence intervals, the Quadratic SVM model using 9 predictors achieved performance metric ranges of [80.4%, 83.8%] for accuracy, [81.9%, 86.2%] for precision, [82.0%, 86.2%] for recall, and [82.5%, 85.8%] for the F1-score.

5. Discussion

In today's world, diabetes is a significant health issue affecting people globally. It occurs when the body struggles with high blood glucose levels due to insulin resistance, a vital hormone for regulating blood sugar. Insulin helps keep the balance in check, and when this balance is disrupted, it can develop into various complications affecting different body parts. Diabetes is influenced by both genetic factors and lifestyle choices, making it a complex health concern that requires ongoing attention to prevent complications. The study analyzed a dataset with 21 factors (excluding labels) to identify the most crucial elements for diagnosing diabetes. The nine key predictor variables pivotal for making an accurate diagnosis were identified. Our innovative AI model, designed for efficiency, demonstrated an impressive classification accuracy of 75.4% using only these nine essential features. This discovery not only provides valuable insights for physicians but also has the potential to streamline their efforts, reducing the workload on healthcare staff and computational resources.

Furthermore, we utilized patient information from Kaggle, a platform known for data science competitions, for specific purposes such as training and testing our models. Unbiased datasets processed in MATLAB highlighted top-performing models, with Quadratic SVM, Coarse Gaussian SVM, and Narrow Neural Networks achieving the highest training accuracy of 76.3%. In a test dataset with some bias, Bilayered Neural Networks led with an accuracy of 74.7%. When considering the average accuracy, precision, recall, and F-1 score, Quadratic SVM consistently emerged as the superior model. To refine our approach, we implemented a feature selection process to rank predictors based on accuracy. Subsequently, the selected predictors underwent retraining to identify the minimal subset that could maintain accuracy comparable to all 21 predictors. This process reduced the parameter count to nine while preserving substantial accuracy. Among the 34 models trained, Quadratic SVM exhibited exceptional performance, boasting average accuracy, precision, recall, and F-1 score of 77.7%. When tested with 21 predictors, Quadratic SVM maintained an average accuracy, precision, recall, and F-1 score of 76.1%, solidifying its position as the most satisfying model. By training with the selected nine predictors, the accuracy reached 75.4%, demonstrating negligible deviation from employing all 21 predictors. Compared to the research that existed before our research focused on optimizing diabetes prediction by employing statistical machine learning models and identifying essential clinical variables. With an impressive classification accuracy of 75.4% using only nine key predictor variables, the study showcased the potential to enhance diagnostic efficiency and alleviate healthcare staff workload. Additionally, among the 34 trained models, Quadratic SVM stood out with superior performance, boasting an average accuracy, precision, recall, and F-1 score of 77.7%. In contrast, the existing research expanded upon this foundation by comparing the performance of various classifiers in predicting type 2 diabetes. The study demonstrated the efficacy of machine learning algorithms in diabetes prediction by achieving high test accuracy ranging from 74.3% to 82.4%. While the neural network model exhibited the highest accuracy and AUC values, it showed lower sensitivity than the decision tree model, which excelled in capturing true positive cases. Despite methodological differences, both studies underscore the potential of machine learning approaches in enhancing diabetes prediction and management. This proposed approach enhances diagnostic efficiency and holds promise for resource optimization in diabetes management, providing a valuable contribution to healthcare practices. This research predicts diabetes without relying on blood glucose or HbA1c. Ongoing efforts to explore alternative methods based on lifestyle patterns, genetics, and non-invasive markers enable readers to prescreen on themselves. Through meticulous analysis, our research has identified the crucial predictors for diagnosing diabetes, enabling the development of an AI model that achieves remarkable accuracy using just nine essential features. This breakthrough streamlines the diagnostic process and offers the potential for quicker and more precise diagnoses, ultimately improving patient outcomes. By implementing this more efficient diagnostic tool, healthcare professionals can alleviate their workload, redirecting time and resources toward providing personalized care and support to patients. Furthermore, our approach holds promise for optimizing resources in diabetes management, as it requires fewer predictors while maintaining comparable accuracy, leading to cost savings and better resource allocation within healthcare settings. Additionally, by exploring alternative methods beyond traditional blood tests, our research enhances accessibility to diabetes screening, allowing individuals to prescreen for risk factors based on lifestyle patterns and genetics, thus empowering proactive health management. While achieving an impressive classification accuracy of 75.4% using only nine essential features is a notable accomplishment, there remains room for improvement in model performance. Further validation efforts by utilizing diverse datasets and external validation with real-world clinical data are necessary to provide a more comprehensive assessment of the model's accuracy and generalizability. Additionally, the clinical relevance of the AI model in predicting diabetes needs to be thoroughly evaluated through clinical validation studies involving healthcare professionals and real patients. These studies will help assess the practical utility of the model in clinical settings and its potential to impact patient outcomes positively. Moreover, it's essential to recognize that diabetes is a multifactorial disease influenced by various genetic, environmental, and lifestyle factors. While the selected predictors capture a subset of these factors, there may be other important determinants of diabetes risk that are not accounted for, complicating the interpretation of the findings.

6. Conclusion

Our research team utilized a Kaggle-derived diabetes dataset encompassing 21 predictors across 253,680 patients. After dividing the dataset into training (67,158 rows) and test (186,522 rows) sets at a 95% to 5% ratio, unbiased training and biased test datasets were randomly selected to ensure equal representation of 0 and 1 cases. Curation processes were applied to establish an unbiased training environment. Machine learning models, such as Quadratic SVM, Coarse Gaussian SVM, and Narrow Neural Networks, achieved the highest training accuracy at 76.3%. However, bias was detected in the test datasets. The Bilayered Neural Network had the highest validation accuracy at 74.7%, but Quadratic SVM emerged as the top performer when averaging accuracy, precision, recall, and F-1 scores. Feature selection employed ranking algorithms like MRMR, Chi^2, reliefF, ANOVA, and Kruskal Wallis. Among the 21 predictors, the 9 selected by Chi^2, including HighBP, HighChol, BMI, HeartDisease, PhysActivity, GenHlth, PhysHlth, DiffWalk, and Age, yielded the highest accuracy at 75.4%.

The research objectives align closely with the broader context of diabetes as a complex and multifactorial health issue. By utilizing statistical machine learning models and focusing on essential clinical variables for diagnosing diabetes, the study aims to address the pressing need for more accurate and efficient diagnostic tools in healthcare. Furthermore, comparing predicted accuracy rates with existing literature findings underscores the importance of benchmarking and advancing current diagnostic approaches. This research not only sheds light on the effectiveness of machine learning algorithms in improving diabetes diagnosis but also contributes valuable insights into the interplay of various clinical and patient characteristics in predicting diabetes. Moreover, the emphasis on model performance, clinical relevance, and ethical considerations reflects a comprehensive approach toward enhancing healthcare practices and patient outcomes in diabetes management. However, it is crucial to acknowledge the limitations and complexities inherent in predicting diabetes accurately, underscoring the need for further validation studies, considering additional influencing factors, and adhering to ethical standards to ensure the robustness and applicability of the findings in clinical settings.

7. Acknowledgements

We acknowledge the Behavioral Risk Factor Surveillance System (BRFSS) for providing the dataset essential to this study. This survey, encompassing responses from over 400,000 Americans, contributes valuable insights into health-related risk behaviors, chronic health conditions, and preventative service utilization. The dataset for 2015, accessible on Kaggle, was generously shared in CSV format by Alex Teboul. The author thanks Satriwitaya School and the Research Institute of Rangsit University for their support and extends special gratitude to the family for their support, encouragement, and backing.

8. References

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://ph04.tci-thaijo.org/index.php/JCST/article/view/327

Ahmed, A. M. (2002). History of diabetes mellitus. *Saudi Medical Journal*, *23*(4), 373-378. https://www.researchgate.net/publication/336666069_History_of_Diabetes_Mellitus

Ahmed, U., Issa, G. F., Khan, M. A., Aftab, S., Khan, M. F., Said, R. A., ... & Ahmad, M. (2022). Prediction of diabetes empowered with fused machine learning. *IEEE Access*, *10*, 8529–8538. https://doi.org/10.1109/access.2022.3142097

Almahdawi, A., Naama, Z. S., & Al-Taie, A. (2022, December 27-28). *Diabetes Prediction Using Machine Learning* [Conference presentation]. 2022 3rd Information Technology To Enhance e-learning and Other Application (IT-ELA), IEEE, Baghdad, Iraq. https://doi.org/10.1109/IT-ELA57378.2022.10107919

Banday, M. Z., Sameer, A. S., & Nissar, S. (2020). Pathophysiology of diabetes: An overview. *Avicenna Journal of Medicine*, *10*(4), 174-188. https://doi.org/10.4103%2Fajm.ajm_53_20

Guan, Z., Li, H., Liu, R., Cai, C., Liu, Y., Li, J., ... & Sheng, B. (2023). Artificial intelligence in diabetes management: advancements, opportunities, and challenges. *Cell Reports Medicine*, *4*(10), Article 101213. https://doi.org/10.1016/j.xcrm.2023.101213

Hart, P. A., Bellin, M. D., Andersen, D. K., Bradley, D., Cruz-Monserrate, Z., Forsmark, C. E., ... & Chari, S. T. (2016). Type 3c (pancreatogenic) diabetes mellitus secondary to chronic pancreatitis and pancreatic cancer. *The Lancet Gastroenterology & Hepatology*, *1*(3), 226-237. https://doi.org/10.1016/s2468-1253(16)30106-6

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

Lu, D., Tao, A., Zeng, T., & Shalaginov, M. (2023, July 19-21). *Machine Learning Techniques for Early Prediction of Diabetes on Multiple Datasets* [Conference presentation]. 2023 3rd International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME), 1-5. Canary Islands, Spain. https://doi.org/10.1109/ICECCME57830.2023.10252198.

Lyngdoh, A., Choudhury, N., & Moulik, S. (2021, March 1-3). *Diabetes Disease Prediction Using Machine Learning Algorithms* [Conference presentation]. 2020 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES), Langkawi Island, Malaysia. https://doi.org/10.1109/IECBES48179.2021.9398759

Manosroi, W., Phimphilai, M., Waisayanand,N., Buranapin, S., Deerochanawong, C., Gunaparn, S., & Phrommintikul, A. (2023). Glycated hemoglobin variability and the risk of cardiovascular events in patients with prediabetes and type 2 diabetes mellitus: A post-hoc analysis of a prospective and multicenter study. *Journal of Diabetes Investigation*, *14*(12), 1391-1400. https://doi.org/10.1111/jdi.14073

Niramitmahapanya, S., Chattieng, P., Nasomphan. & T., Sathirakul, K. (2023). Effects of dietary supplementation on progression to type 2 diabetes in subjects with prediabetes: a single center randomized double-blind placebo-controlled trial. *Annals of Clinical Endocrinology and Metabolism*, 7, 001-007.

Panda, N. R., Mohanty, J. N., Bhuyan, R., Raut, P. K., & Manulata. (2024). Exploring machine learning approaches for early diabetes risk prediction: A comprehensive examination of health indicators and models. *Journal of Associated Medical Sciences*, *57*(3), 155–165. Retrieved from https://he01.tci-thaijo.org/index.php/bulletinAMS/article/view/271446

Popoviciu, M. S., Kaka, N., Sethi, Y., Patel, N., Chopra, H., & Cavalu, S. (2023). Type 1 Diabetes Mellitus and Autoimmune Diseases: A Critical Review of the Association and the Application of Personalized Medicine. *Journal of Personalized Medicine*, *13*(3), Article 422. https://doi.org/10.3390/jpm13030422

Sims, E. K., Carr, A. L., Oram, R. A., DiMeglio, L. A., & Evans-Molina, C. (2021). 100 years of insulin: celebrating the past, present and future of diabetes therapy. *Nature Medicine*, *27*(7), 1154-1164. https://doi.org/10.1038/s41591-021-01418-2

Sugandh, F. N. U., Chandio, M., Raveena, F. N. U., Kumar, L., Karishma, F. N. U., Khuwaja, S., ... & Kumar, S. (2023). Advances in the management of diabetes mellitus: a focus on personalized medicine. *Cureus*, *15*(8), Article e43697. https://doi.org/10.7759%2Fcureus.43697

Teboul, A., (2022). Diabetes Health Indicators Dataset. Kaggle, Retrieved October 15, 2023 from https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset/data

Xie ZiDian, X. Z., Nikolayeva, O., Luo JieBo, L. J., & Li DongMei, L. D. (2019). Building risk prediction models for type 2 diabetes using machine learning techniques. *Preventing Chronic Disease*, *19*(16), Article e130. https://doi.org/10.5888/pcd16.190109