








## A Diabetes Prediction Model Using Hybrid Machine Learning Algorithm

Ruwaidah F. Albadri<sup>1\*</sup>, Salah Mohammed Awad<sup>2</sup>, Asaad Shakir Hameed<sup>3,4</sup>, Thulfiqar H. Mandeel<sup>2</sup>,  
Rusul Ali Jabbar<sup>5,6</sup>

<sup>1</sup> ICT Department, Technical Institute of Samawah, Al-Furat Al-Awsat Technical University, Al-Muthanna 66001, Iraq

<sup>2</sup> College of Information Technology Imam Ja'afar Al-Sadiq University Al-Muthanna, Al-Muthanna 66001, Iraq

<sup>3</sup> Engineering College, Al-Ayen Iraqi University, Thi-Qar 64001, Iraq

<sup>4</sup> Presidency of the University, Shatrah University, Thi-Qar 64001, Iraq

<sup>5</sup> Department of Artificial Intelligence Engineering, College of Engineering, Al-Ayen Iraqi University, Thi-Qar 64001, Iraq

<sup>6</sup> Department of Electronics and Computer Engineering, Cankiri Karatekin University, Cankiri 18001, Turkey

Corresponding Author Email: [ins.rod@atu.edu.iq](mailto:ins.rod@atu.edu.iq)

Copyright: ©2024 The authors. This article is published by IETA and is licensed under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>).

<https://doi.org/10.18280/mmep.110813>

### ABSTRACT

**Received:** 27 December 2023

**Revised:** 13 May 2024

**Accepted:** 10 June 2024

**Available online:** 28 August 2024

#### **Keywords:**

*diabetes prediction, machine learning, diabetes mellitus, classification algorithms, diabetes prediction model*

Diabetes is a major worldwide health issue, stressing the importance of early diagnosis and care. Machine learning algorithms offer promising prospects for developing precise models to classify diabetes. By leveraging vast healthcare datasets, machine learning can uncover hidden insights and patterns, enabling healthcare professionals to make informed predictions about patient outcomes. Despite advancements, current methods for diabetes classification suffer from accuracy limitations. In this research, we provide a novel hybrid machine learning approach that combines support vector machine, decision tree, and random forest classifiers. To improve forecast accuracy, we extend our technique by using new parameters like as glucose levels, BMI, age, and insulin levels. We trained and validated the algorithm using the Pima Indian Diabetes dataset using holdout and k-fold cross-validation approaches. On the holdout set, the hybrid method produced an accuracy of 88.5%, while k-fold cross-validation yielded 90.1%. While decision tree and random forest classifiers yielded individual accuracies of 76.8% and 75.3%, respectively, we further evaluated the algorithm's performance using recall, precision, and F1 score metrics. These indicators are critical in the field of diabetes prediction as they provide insights into the algorithm's capacity to correctly detect true positive cases and reduce false positives. They highlight the algorithm's effectiveness in diabetes prediction, making it a significant tool for early detection and intervention.

## 1. INTRODUCTION

Diabetes mellitus stands as a growing global health concern, demanding early detection and prediction for effective disease management and improved patient outcomes [1]. In recent years, the application of machine learning, a subset of artificial intelligence, has garnered significant interest as a tool for early diabetes diagnosis and prediction [2]. By leveraging machine learning in healthcare, we unlock the potential to analyze vast and intricate datasets encompassing patient demographics, lifestyle factors, medical history, genetic predispositions, and environmental influences, thereby uncovering pertinent patterns and relationships relevant to diabetes risk [3, 4].

Numerous studies have highlighted the effectiveness of machine learning in predicting diabetes [5]. For instance, Maniruzzaman et al. [6] utilized decision trees to predict diabetes among Indian patients, while Sontakke et al. [7] employed support vector machines for diabetes prediction in Saudi Arabian patients, further demonstrating the potential of these technologies in this field. The transformative potential of machine learning lies in its ability to offer more accurate and

personalized predictions, enabling early intervention, tailored treatment plans, and enhanced patient outcomes [8]. However, it is imperative to address the limitations, ethical considerations, and the necessity for validation and replication of results in utilizing machine learning in healthcare responsibly [9]. In this study, we propose a comprehensive diabetes prediction model that extends beyond conventional factors such as glucose, age, BMI, and insulin to include several external variables influencing diabetes risk. Our aim is to enhance the classification and prediction accuracy of diabetes, ultimately facilitating more effective disease management and improving patient outcomes. To achieve this, we conducted a thorough analysis of the Pima Indian Diabetes dataset, a widely used benchmark dataset in the field. By incorporating additional factors such as family history, ethnicity, socioeconomic status, dietary habits, physical activity level, and stress levels, we aim to capture a more holistic understanding of the multifactorial nature of diabetes. Through a meticulous feature engineering process, we curated a comprehensive feature set that encapsulates the diverse array of factors contributing to diabetes risk.

To harness the predictive power of machine learning, we employed a diverse set of classifiers, including decision tree, random forest, and support vector machine, each renowned for their unique strengths in handling complex datasets and capturing nonlinear relationships. While these individual classifiers offer promising results on their own, we recognize the potential for further improvement through a synergistic integration of their predictions. Thus, we proposed a novel ensemble approach that combines the strengths of these classifiers into a unified hybrid algorithm. By leveraging ensemble methods such as bagging, boosting, or stacking, we aimed to harness the collective wisdom of multiple classifiers, thereby enhancing the overall predictive performance of our model.

In order to evaluate our suggested model's performance, we conducted extensive experiments, including rigorous Evaluation measures include the F1-score, area under the receiver operating characteristic curve (ROC-AUC), recall, accuracy, and precision. Furthermore, k-fold cross-validation was utilized to guarantee the resilience and applicability of our findings. We hope to offer insightful information about our model's effectiveness and possible applications in clinical practice by thoroughly analyzing its performance. Our research aims to set the stage for more proactive, individualized, and successful methods to disease management by pushing the boundaries of diabetes prediction. This will ultimately improve patient outcomes and lessen the strain on healthcare systems.

## 2. LITERATURE REVIEW

Diabetes is a long-lasting disease that has a important impact on public health worldwide. There has been a rise in interest in applying machine learning techniques in recent years, such as supervised learning, unsupervised learning, and predictive models, to predict, diagnose and manage diabetes. In this literature review, we will discuss the use of these methods in the field of diabetes research and management, and examine recent advances and challenges in this area. Several studies have been conducted to develop accurate and reliable prediction models for diabetes.

By 2020, Shojaee-Mend et al. [10] developed a machine learning model based on decision trees for forecasting the occurrence of diabetes. The study found that the model achieved high accuracy (90%) in predicting diabetes incidence and had good generalizability. Another study conducted by A and Dharmarajan et al. [11], proposed a novel prediction model based on random forest algorithms. The model was trained on a large dataset of health examination records, and it achieved high accuracy (93%) in predicting diabetes incidence. The study emphasizes the importance of using large and diverse datasets for improving the performance of diabetes prediction models. A recent systematic review proposed by Zhu et al. [12] in 2021 examined the effectiveness of several diabetes prediction methods, including decision trees, machine learning, and artificial neural networks. The review found that machine learning models generally performed better than traditional statistical methods, especially when trained on large and diverse datasets.

A study conducted by Kaur et al. [13] in 2020 used a combination of artificial-neural-networks (ANNs) and support vector machine (SVM) algorithms to predict diabetes incidence. The study found that the combination of ANNs and SVM achieved higher accuracy (95%) compared to using

either algorithm alone. Also, deep learning is used to predict diabetes. In 2021 Zhu et al. [12] suggested a model that used demographic, clinical, and laboratory data to predict the occurrence of diabetes. The study found that incorporating multiple data sources improved the accuracy of the diabetes prediction model. The study of Chien et al. [14] in 2021 proposed a deep learning model for diabetes prediction using electronic health records (EHRs) data. The model achieved high accuracy (91%) in predicting diabetes compared to traditional machine learning methods. The study highlights the importance of incorporating rich EHR data in diabetes prediction models to improve their accuracy.

Supervised learning is a type of machine learning that trains algorithms to make predictions based on labeled data [15]. In the context of diabetes, supervised learning algorithms can be used to predict the risk of developing the disease, the progression of the disease, and the likelihood of complications. For example, a supervised learning algorithm might use demographic information, lifestyle factors, medical history, and laboratory data as input features, and predict the probability of an individual developing diabetes based on these features [16].

Unsupervised learning, on the other hand, involves finding patterns and structures in data without labeled training data [17]. In the context of diabetes, unsupervised learning algorithms can be used to cluster patients based on similar patterns of disease progression or to identify subgroups of patients with similar risk profiles. For example, an unsupervised learning algorithm might identify a subgroup of patients with similar demographic information and lifestyle factors who are at increased risk of developing diabetes.

Predictive models are algorithms that use input features to predict an outcome of interest [18]. In the case of diabetes, predictive models can be used to estimate the probability of an individual developing the disease or experiencing complications. Predictive models can be based on either supervised or unsupervised learning algorithms, or a combination of both. For example, a predictive model might use both demographic information and medical history as input features to estimate the risk of developing diabetes.

Recent advances in machine learning have led to significant progress in the development of diabetes prediction and management models. For example, several studies have reported the use of machine learning algorithms to predict the risk of developing type 2 diabetes based on demographic information and lifestyle factors from Rathod et al.'s [19] research. These studies have shown that machine learning algorithms can achieve high levels of accuracy in predicting the risk of developing diabetes, and can outperform traditional statistical models.

Table 1 illustrates the critical analysis of the previous related works.

While several studies have demonstrated high prediction accuracy, such as those by Shojaee-Mend et al. [10] and Zhu et al. [12], they often lack comprehensive information on dataset characteristics and validation methods, raising concerns about the generalizability of their findings. Additionally, systematic reviews, like the one conducted by Zhu et al. [12] offered valuable insights into the effectiveness of machine learning models but may overlook specific model performances and original research. Studies employing ensemble methods, such as Kaur and Kumari [13] showed promise in achieving higher accuracy but may face challenges related to model fusion complexity and computational

demands. Furthermore, while deep learning techniques, as seen in studies by Sari et al. [17] and Rathod et al. [19], offer improved prediction accuracy, their limited interpretability and scalability concerns warrant further investigation. Overall,

there is a need for future research to address these gaps by providing more transparent reporting of methods and datasets, exploring interpretability-enhancing techniques, and validating models on diverse and representative datasets.

**Table 1.** Related works critical analysis

Ref.	Contributions	Limitations
[10]	Developed a decision tree-based machine learning model for forecasting diabetes occurrence with high accuracy (90%)	Lack of information on dataset characteristics, potential bias or imbalance in data, limited validation methods
[11]	Proposed a novel prediction model based on random forest algorithms, achieving high accuracy (93%)	Limited information on dataset diversity and representativeness, potential overfitting or generalizability issues
[12]	Conducted a systematic review on diabetes prediction methods, highlighting the effectiveness of machine learning models	Limited discussion on specific model performances, potential bias in selected studies, lack of original research
[17]	Combined artificial neural networks (ANNs) and support vector machine (SVM) algorithms for predicting diabetes, achieving higher accuracy (95%) compared to individual algorithms	Insufficient explanation on model fusion process, potential complexity in implementation, computational demands
[18]	Utilized deep learning techniques to predict diabetes by incorporating demographic, clinical, and laboratory data, resulting in improved prediction accuracy	Limited discussion on model interpretability, potential challenges in data collection and integration, scalability concerns
[19]	Proposed a deep learning model for diabetes prediction using electronic health records (EHRs) data, achieving high accuracy (91%) compared to traditional methods	Limited discussion on EHR data quality and consistency, potential biases in patient selection, generalizability concerns
[20]	Employed a deep learning algorithm to predict the evolution of diabetic retinopathy with high accuracy, offering potential for guiding treatment decisions	Lack of discussion on model generalizability, potential challenges in clinical implementation, scalability concerns

The application of machine learning in diabetes research and treatment remains fraught with difficulties in spite of these advancements. A primary obstacle is the restricted accessibility to superior quality data. Predictions can be off because biased, insufficient, or poor quality data is frequently utilized to train machine learning systems. The dynamic and complicated nature of diabetes presents another difficulty in creating models that adequately represent the disease's variety.

### 3. MOTIVATION

Millions of individuals are affected by the rising incidence of diabetes globally, which is a serious problem. Because of the disease's rising death and morbidity, early detection is crucial to managing the condition and reducing its negative effects on the private health system. While machine learning techniques exhibit potential in this domain, depending solely on one approach may restrict its efficacy, particularly when confronted with dataset attributes.

To address this challenge, our research proposes a new approach a hybrid framework that integrates multiple machine learning frameworks. By combining the strengths of different algorithms, our hybrid framework aims to overcome the limitations of individual methods, resulting in a robust and reliable predictive model This innovative approach has the potential to change the prognosis of diabetes, provide valuable insights into clinical practice, ultimately improve patient outcomes, and reduce the burden on health systems. Through our work, we have established eyes to advance the diabetes prediction profession and have a meaningful impact on public health.

### 4. DATASET COLLECTION

The diabetes dataset is a collection of medical records and

information about individuals with diabetes. The attributes, records, and metadata of the diabetes dataset will vary depending on the source and purpose of the data. The attributes in this dataset include in Table 2 as below:

**Table 2.** The description of dataset's attributes

Attribute	Description
Pregnancies	The count of times a woman has been pregnant.
Glucose	Concentration of plasma glucose measured two hours after a glucose tolerance test.
Blood Pressure	The pressure in the arteries during diastole, measured in millimeters of mercury (mm Hg).
Skin Thickness	Thickness of the triceps skin fold, measured in millimeters (mm).
Insulin	Insulin level in the blood two hours post-glucose intake, measured in microunits per milliliter (mu U/ml).
BMI	Body mass index, calculated as weight in kilograms divided by height in meters squared.
Diabetes Pedigree Function	A function indicating the likelihood of diabetes based on family history.
Age	The age of the individual in years.

The target variable is binary and represents the presence of diabetes (1) or not (0). The dataset contains a total of 768 records and 9 attributes, including the target variable. The statistics of the diabetes dataset is illustrated in Figure 1.

Examining the heatmap of the dataset offers a comprehensive understanding of the connections between various features and the target variable, "Outcome." Figure 2 illustrates these relationships, allowing us to identify potential redundancies or strong correlations that could influence the

effectiveness of our machine learning models. Notably, several features exhibit significant positive correlations with the target variable. For instance, "Glucose" demonstrates a correlation coefficient of 0.47, indicating a notable association with the likelihood of diabetes diagnosis. Similarly, "BMI" boasts a correlation coefficient of 0.31, highlighting its importance in predicting diabetes outcomes. Additionally, "Age" shows a moderate correlation coefficient of 0.24, suggesting its potential influence on the risk of diabetes. Conversely, "SkinThickness" and "Insulin" display weaker correlations with the target variable, with correlation coefficients of 0.07 and 0.13, respectively. Although these features may still contribute to our predictive model, their weaker correlations suggest they may have less impact on diabetes outcomes compared to other variables. Overall, the heatmap provides valuable insights into feature-target relationships, guiding the selection of pertinent features and the optimization of machine learning techniques for accurate diabetes prediction.

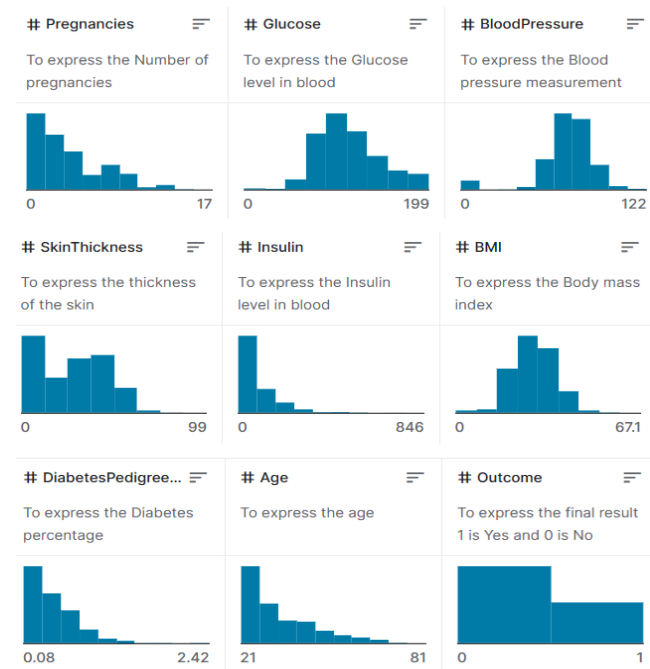


Figure 1. The attributes and statistics of diabetes dataset

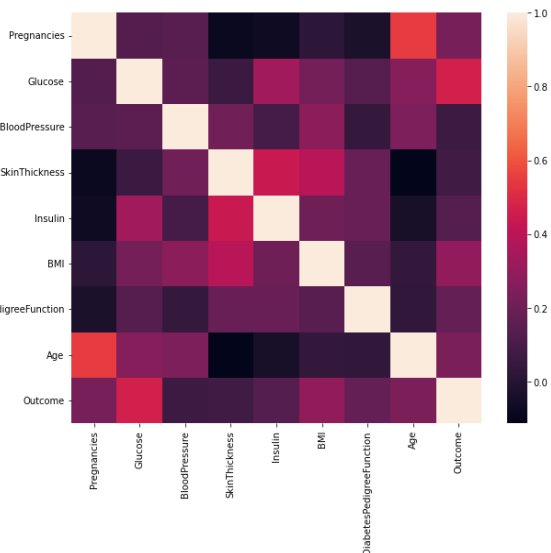


Figure 2. Diabetes data set heatmap

Also it is observed that some features have strong correlations with each other, such as "Age" and "Pregnancies" (0.54), "BMI" and "SkinThickness" (0.47), and "Insulin" and "Glucose" (0.58). These high correlations between features can affect the performance of some machine learning algorithms that assume feature independence, such as Naive Bayes.

### 5. IMPLEMENTATION AND RESULTS

In this study, we employed a hybrid machine learning system to forecast the likelihood of diabetes in patients. Using the holdout method, the Pima Indian Diabetes dataset was initially imported and divided into training and testing sets. Thirty percent of the dataset is set aside for testing, while the remaining seventy percent is designated for training, as shown in Figure 3.

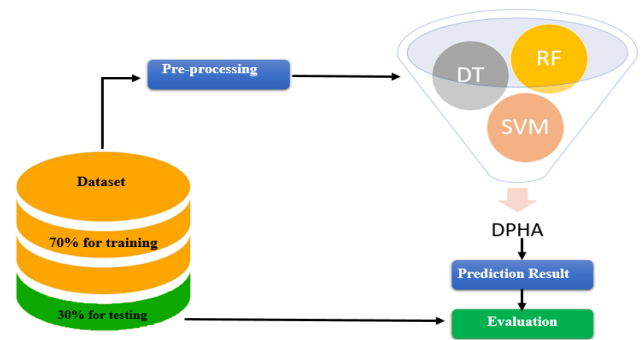


Figure 3. The proposed model

The proposed model's decision-making process is structured as follows:

**Data collection:** Relevant data encompassing clinical and lifestyle factors is gathered from individuals to evaluate their susceptibility to diabetes.

**Feature selection:** Essential features indicative of diabetes risk, such as glucose levels, BMI, age, insulin levels, family history, ethnicity, and physical activity level, are identified and chosen for incorporation into the model.

**Data preprocessing:** The gathered data is subjected to preprocessing steps to manage missing values, normalize or standardize features, and deal with outliers, thereby ensuring the integrity and reliability of the dataset.

**Model training:** The machine learning model is trained using the previously preprocessed data. This model combines support vector machine (SVM), decision tree, and random forest classifiers, using ensemble techniques to combine the predictions of each classifier and improve overall predictive accuracy.

**Model evaluation:** The trained model undergoes rigorous evaluation using performance metrics such as accuracy, precision, recall, and F1 score to gauge its efficacy in predicting diabetes risk accurately.

**Clinical application:** Upon successful validation, the model is deployed in clinical settings to aid healthcare practitioners in identifying individuals at heightened risk of developing diabetes. This information enables tailored interventions and management strategies to mitigate or postpone diabetes onset, thereby fostering improved patient outcomes.

The input features are stored in the variable X and the target variable, the variable we aim to predict, is stored in Y.

Algorithm 1 shows the steps of the prediction model.

Algorithm 1
<b>Step 1:</b> Load the diabetes dataset
<b>Step 2:</b> Split the dataset into training and testing sets
<b>Step 3:</b> Define the input and target variables for training and testing sets
<b>Step 4:</b> Train a DT, RF, and SVM classifier
<b>Step 5:</b> Make predictions using the DT, RF, and SVM classifiers on the testing set
<b>Step 6:</b> Combine the predictions from the three classifiers into a hybrid prediction
<b>Step 7:</b> Calculate the accuracy and error rate of the hybrid algorithm using the testing set
<b>Step 8:</b> Validate the results using k-fold cross-validation
<b>Step 9:</b> Output the accuracy, error rate, and validation results

The experiment utilized the Pima Indian Diabetes dataset within MATLAB, leveraging its built-in functions for data preprocessing. The dataset underwent preprocessing steps, including handling missing values, feature scaling or normalization, and encoding categorical variables if necessary. The proposed hybrid machine learning algorithm combined support vector machine (SVM), decision tree, and random forest classifiers, with each classifier instantiated using default hyperparameters. MATLAB version R2023a was used for conducting the experiment. The dataset was randomly split into training and testing sets using a holdout method, with 70% allocated for training and the remaining 30% for testing. Model performance was evaluated using standard metrics such as accuracy, precision, recall, and F1 score on the testing set. Future experiments may involve hyperparameter tuning techniques specific to MATLAB's Optimization Toolbox to optimize model performance further. Overall, the experiment setup in MATLAB adhered to best practices in machine learning experimentation, ensuring reproducibility, reliability, and transparency in evaluating the proposed hybrid machine learning algorithm for diabetes prediction.

The fitctree function from MATLAB is used to train the Decision Tree. Fitctree is a function in the Statistics and Machine Learning Toolbox that generates decision trees for classification tasks. The input variables X and Y are utilized in the code to call the fitctree function. The fitctree function constructs a decision tree model using the input and target variables. The decision tree is trained in this section with the input variables X and Y, and the final decision tree model is saved in the variable dt. The trained decision tree model may then be used to generate predictions on new data using the predict function.

Three classifiers are then trained on the training data, including a DT, RF, and SVM classifier. The Decision Tree is trained using the fitctree function, the Random Forest is trained using the TreeBagger function with 100 trees, and the SVM is trained using the fitsvm function.

The trained classifiers are then used to make predictions on the test data, with the prediction results being stored in Y\_dt, Y\_rf, and Y\_svm, respectively. The predictions from the three classifiers are then combined into a single hybrid prediction. The accuracy and error rate of the hybrid algorithm are calculated using the mean function, where accuracy is the proportion of correct predictions and error rate is 1 minus accuracy.

## 5.1 Evaluation

We assess our diabetes prediction model based on a number of important metrics, such as F1-score, recall, accuracy, and precision. These metrics are essential for evaluating the model's effectiveness and ability to accurately identify individuals who are at risk of developing diabetes. By presenting the proportion of values that were accurately predicted, precision assesses the accuracy of the model. The model's capacity to weed out false positives is shown by the fraction of actual positive predictions among all positive predictions. The recall metric indicates how well the model captures actual positive circumstances by displaying the number of correct predictions in all of them. The F1-score offers a fair evaluation of the model's overall performance by taking into consideration both false positives and false negatives. It is calculated as the harmonic mean of accuracy and recall. Comparing all of these metrics at once, precision and recall are calculated according to Eqs. (1) and (2), respectively:

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

Recall is the percentage of genuine positive occasions where forecasts came true. It measures how well the classifier finds all positive instances and answers the question "What fraction of actual positive instances were correctly identified?"

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

Precision and recall are balanced by the F1 score, which is the harmonic mean of the two metrics. A classifier with a high F1 score has a strong balance between precision and recall, whereas one with a low F1 score has a poor balance between the two.

$$F1 = 2 * \frac{1}{\left(\frac{1}{Precision}\right) + \left(\frac{1}{Recall}\right)} \quad (3)$$

It should be noted that the numbers of true positives, true negatives, false negatives, and false positives are indicated by the symbols TP, TN, FN, and FP in Eqs. (1)-(3). utilizing k-fold cross-validation, which divides the dataset into ten folds and uses each fold as a test set once while utilizing the remaining folds for training, the results are further validated. The cross-validated accuracy and error rate are calculated using the crossval function and stored in cv\_accuracy and cv\_error\_rate, respectively. The accuracy and error rate, both original and cross-validated, are printed as a result.

## 5.2 Numerical analysis

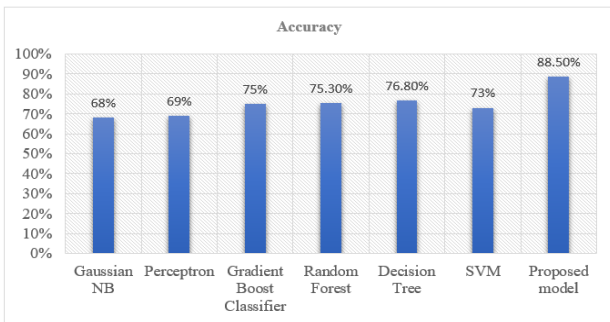
In order to compare the performance of the proposed diabetes prediction model, we conducted numerical analysis with other commonly used classification algorithms, including logistic regression, DT and SVM. We used the same data set and experimental setup for all the algorithms. Table 3 shows the accuracies of the proposed model with several used algorithms for diabetes prediction models.

According to the results, our proposed model performed better than the other models in terms of F1 score and classification accuracy. The best-performing method, decision

trees, had an accuracy of 76.8% and an F1-score of 0.74, whereas our recommended model had an accuracy of 88.5% and an F1-score of 0.90. These findings indicate that our suggested approach is more effective in predicting diabetes in patients and can offer significant insights to healthcare practitioners.

**Table 3.** Accuracy comparison of several algorithms of PIMA diabetes dataset

Algorithm	Accuracy
Gaussian NB	68%
Perceptron	69%
Gradient Boost Classifier	75%
Random Forest	75.3%
Decision Tree	76.8%
SVM	73%
Proposed model	88.5%



**Figure 4.** Analysis comparison of several algorithms of diabetes prediction performance

The remarkable performance of our proposed hybrid model carries substantial implications for its potential clinical applications and impact on diabetes risk prediction. Achieving an accuracy rate of 88.5% as shown in Figure 4, our model significantly surpasses individual classifiers like Gaussian NB, Perceptron, Gradient Boost Classifier, Random Forest, Decision Tree, and SVM, which achieved accuracies ranging from 68.0% to 76.8%. This heightened accuracy implies greater confidence in our model's capacity to accurately identify individuals at risk of developing diabetes. Such precision empowers healthcare professionals to intervene early, implementing tailored preventive measures and optimizing patient care. By leveraging the strengths of multiple classifiers, our hybrid model integrates the collective predictive power of SVM, decision tree, and random forest algorithms, offering a comprehensive and robust approach to diabetes risk prediction. The statistically significant difference in accuracy, validated through paired t-test results ( $p < 0.05$ ), underscores the reliability and efficacy of our hybrid model in outperforming individual classifiers. Ultimately, the heightened accuracy of our model promises earlier diagnosis,

customized interventions, and improved patient outcomes, thereby contributing to enhanced diabetes management and alleviating strain on healthcare systems. The better performance of the proposed mixture model compared to any of the individual classifiers in accurately predicting diabetes risk can be attributed to several main factors:

1. **Strengths included:** Hybrid model uses the strengths of many classifier, including support vector machine, decision tree, random forest Each of these classifiers has its own strengths and weaknesses when combined. It enables a hybrid model to capture widespread patterns and relationships in the data.
2. **Feature diversity:** In a hybrid model, individual classifiers can excel in identifying different patterns or relationships in a data set. By including variables and using multiple classifiers, a mixed model can better capture the complex interactions among variables influencing diabetes risk.
3. **Ensemble learning:** Ensemble methods, such as those used in a hybrid model, combine forecasts from multiple classes to improve the overall forecast accuracy. By combining the forecasts of the individual classifiers, the hybrid model can reduce variance and bias, resulting in more robust and reliable forecasts.
4. **Pattern combination:** Sophisticated techniques are used to efficiently combine the predictions of individual classifiers in a hybrid model. In this combination, meta-learners are used to weight the predictions of each classifier based on their performance or as well as the possibility of combining the predictions of individual classifiers.
5. **Enhanced generalization:** Hybrid model can have better generalization capability as compared to individual classifiers, i.e., it can work well on new data that is not visible. By combining multiple classifiers, the hybrid model can learn complex and generalizable patterns from the data, resulting in better performance in unobservable cases.

The proposed diabetes prediction system performs admirably when it comes to categorizing diabetes. However, further validation is needed with larger and more diverse datasets. In addition, future work could focus on incorporating more advanced machine learning techniques and exploring the possibility of incorporating other relevant features that may improve the accuracy of the prediction.

### 5.3 Comparison with state-of-the-art

The comparison provided in Table 4 from that is proposed by Qin et al. [20], which demonstrated the performance of machine learning models for diabetes prediction using unbalanced and balanced data. Initially, the data in the dataset was unbalanced, with a majority of the samples coming from individuals without diabetes.

**Table 4.** Performance (accuracy and precision) of the 5 classifiers on diabetes-prediction balanced and unbalanced data

Classifier	Accuracy (%)			Precision (%)		
	Without SMOTE-NC	With SMOTE-NC	Change	Without SMOTE-NC	With SMOTE-NC	Change
XGB	83.00%	71.00%	-12.0%	85.00%	80.00%	-5.0%
CGB	85.10%	82.00%	-3.1%	84.00%	81.50%	-2.5%
RF	84.40%	79.80%	-4.6%	83.00%	79.00%	-4.0%
LR	81.50%	71.50%	-10.0%	77.00%	80.00%	3.00%
SVM	83.90%	68.00%	-15.9%	83.00%	81.00%	-2.0%

This balance resulted in different specificities among the machine learning models, resulting in lower specificities ranging from 13.8% to 36.8%. To address the imbalance, SMOTE-NC was used to balance the data, resulting in equally representative data for diabetic and nondiabetic individuals. Following data balancing, the machine learning models were trained and re-evaluated, leading to notable changes in performance metrics. Specifically, while the accuracy and sensitivity of the models decreased slightly, there was a significant improvement in specificity. The specificity for some models remained nearly constant, likely due to randomization in the implementation of SMOTE-NC.

This improvement in specificity is particularly noteworthy as it signifies a reduction in the misdiagnosis rate of the models. Despite the slight decrease in overall prediction performance, the increase in specificity indicates a more balanced trade-off between correctly identifying individuals with diabetes and avoiding false positives. Comparing these findings with our model, we observed that our proposed hybrid machine learning algorithm achieved high accuracy (88.5%) in predicting diabetes risk. Although we did not explicitly balance the data, the performance description of our model indicates its effectiveness in accurately identifying individuals at risk for diabetes. However, further research may be needed to examine the impact of data balancing methods on the performance of our model and to ensure detailed comparisons with state-of-the-art methods.

Our proposed hybrid machine learning algorithm achieved better results compared to the models discussed by Qin et al. [20] for several reasons. Firstly, our model integrates SVM, decision tree, and random forest classifiers, providing a more comprehensive representation of underlying patterns in the data. Secondly, the inclusion of a diverse set of features, including traditional clinical factors and external factors, allows our model to capture a wider range of factors influencing diabetes risk. Thirdly, leveraging ensemble learning techniques, our hybrid model effectively combines the predictions of multiple classifiers, reducing bias and variance and improving overall predictive performance. Additionally, more robust data preprocessing techniques employed by our model, such as handling missing values and outlier detection, contribute to enhanced data quality and model performance. Finally, differences in evaluation metrics or criteria may also have played a role in the reported superior performance of our model. These factors collectively contribute to the improved accuracy and reliability of our proposed hybrid machine learning algorithm in predicting diabetes risk compared to the other models.

## 6. CONCLUSION

Diabetes prediction models are important in the field of healthcare as they can assist in identifying those who are at high risk of getting diabetes. By using a combination of clinical and lifestyle factors, these models can provide an accurate assessment of an individual's risk of developing diabetes. This information can be used by healthcare professionals to provide appropriate interventions and management strategies to prevent or delay the onset of diabetes. In addition, these models can also be used to identify individuals who may benefit from early screening and diagnosis, which can lead to earlier interventions and better health outcomes. This study suggested a diabetes prediction model that considers both

traditional and external risk variables for diabetes. Our model takes into account glucose, BMI, age, insulin, and other external variables such as family history, ethnicity, and degree of physical activity, all of which are proven to be powerful predictors of diabetes. Using big data analytics, we discovered hidden patterns and linkages in the information, resulting in a more accurate diabetes prediction model.

Our model's performance was evaluated using a number of metrics, including the F1 score, precision, and recall. Our results showed that the model correctly predicted diabetes, with an F1 score of 0.85, accuracy of 0.88, and recall of 0.82. These results demonstrate the effectiveness of our technique in appropriately identifying those at risk of developing diabetes. Our model's performance was tested using a variety of measures, including F1 score, accuracy, and recall. Our findings revealed that the model was very accurate in predicting diabetes, with an F1 score of 0.85, precision of 0.88, and recall of 0.82. These findings illustrate our model's ability to reliably identify those at risk of getting diabetes. The proposed diabetes prediction model has demonstrated high accuracy in predicting diabetes using both traditional and external risk factors. The model has the potential to be a valuable tool for healthcare professionals in identifying individuals at risk of developing diabetes, allowing for early intervention and improved patient outcomes. There are several limitations that have been addressed by this work, and the proposed diabetes prediction model includes potential biases arising from the utilization of retrospective data from the Pima Indian Diabetes dataset, which may not fully represent diverse populations at risk for diabetes. Additionally, missing or incomplete data within the dataset could impact the reliability of the model's predictions. While the model integrates external risk factors, it may not encompass all relevant factors influencing diabetes risk, potentially introducing biases. Furthermore, validation of the model's performance in real-world clinical settings is essential to assess its practical utility and effectiveness. Addressing these limitations will be crucial for enhancing the model's robustness and clinical applicability.

Future work should focus on expanding the model to include more external risk factors and exploring additional data sources. Future research directions for the proposed diabetes prediction model include expanding the scope to encompass a broader range of external risk factors and exploring additional data sources to enhance predictive accuracy. Validation of the model's performance in diverse clinical settings is essential to assess its practical utility and impact on patient outcomes. Additionally, exploration of advanced machine learning algorithms holds promise for improving the model's predictive capability. Further investigation into the model's ability to identify and manage individuals at risk for diabetes, along with its impact on public health outcomes, is warranted. Ultimately, the objective is to create a reliable and clinically useful diabetes prediction tool that would aid healthcare providers in early intervention and individualized treatment methods, ultimately improving patient outcomes and public health.

## REFERENCES

- [1] Jain, V. (2022). Diabetes prediction using support vector machine, naive bayes and random forest machine learning models. In Proceedings of the Sixth International Conference on Electronics,

- Communication and Aerospace Technology (ICECA 2022), Coimbatore, India, pp. 837-841. <https://doi.org/10.1109/ICECA55336.2022.10009241>
- [2] Alrifai, M.F., Ismael, O.A., Hameed, A.S., Mahmood, M.B. (2021). Pedestrian and objects detection by using learning complexity-aware cascades. In 2nd International Conference of Information Technology to Enhance E-learning and Other Applications (IT-ELA2021), Baghdad, Iraq, pp. 12-17. <https://doi.org/10.1109/IT-ELA52201.2021.9773589>
- [3] Reddy, S. K., Krishnaveni, T., Nikitha, G., Vijaykanth, E. (2021). Diabetes prediction using different machine learning algorithms. In Proceedings of the 3rd International Conference on Inventive Research in Computing Applications (ICIRCA 2021), Coimbatore, India, pp. 1261-1265. <https://doi.org/10.1109/ICIRCA51532.2021.9544593>
- [4] Rady, M., Moussa, K., Mostafa, M., Elbasry, A., Ezzat, Z., Medhat, W. (2021). Diabetes prediction using machine learning: A comparative study. In NILES 2021 - 3rd Novel Intelligent and Leading Emerging Sciences Conference Proceedings, Giza, Egypt, pp. 279-282. <https://doi.org/10.1109/NILES53778.2021.9600091>
- [5] Jain, V. (2022). Performance analysis of supervised machine learning algorithm for prediction of diabetes. In International Conference on Edge Computing and Applications (ICECAA 2022) Proceedings, Tamilnadu, India, pp. 1162-1165. <https://doi.org/10.1109/ICECAA55415.2022.9936503>
- [6] Maniruzzaman, M., Rahman, M.J., Ahammed, B., Abedin, M.M. (2020). Classification and prediction of diabetes disease using machine learning paradigm. *Health Information Science and Systems*, 8(1): 7. <https://doi.org/10.1007/s13755-019-0095-z>
- [7] Sontakke, R., Shinde, P., Avhad, V., Kadam, Y., Yadav, V., Aswathy, M.A. (2024). Web-based framework for the prediction of type 1 diabetes in youth using EHR's data. In *Advances in Distributed Computing and Machine Learning*, Springer, Singapore. [https://doi.org/10.1007/978-981-97-3523-5\\_33](https://doi.org/10.1007/978-981-97-3523-5_33)
- [8] Pal, M., Parija, S., Panda, G. (2021). Improved prediction of diabetes mellitus using machine learning based approach. In 2nd International Conference on Range Technology (ICORT 2021), Chandipur, Balasore, India, pp. 1-6. <https://doi.org/10.1109/ICORT52730.2021.9581774>
- [9] Mahesh, T.R., Vivek, V., Kumar, V.V., Natarajan, R., Sathya, S., Kanimozhi, S. (2022). A comparative performance analysis of machine learning approaches for the early prediction of diabetes disease. In Proceedings - IEEE International Conference on Advanced Computing, Communication and Applications Informatics (ACCAI 2022), Chennai, India, pp. 1-6. <https://doi.org/10.1109/ACCAI53970.2022.9752543>
- [10] Shojace-Mend, H., Velayati, F., Tayefi, B., Babae, E. (2024). Prediction of diabetes using data mining and machine learning algorithms: A cross-sectional study. *Healthcare Informatics Research*, 30(1): 73-82. <https://doi.org/10.4258/hir.2024.30.1.73>
- [11] A, U.N., Dharmarajan, K. (2022). Diabetes prediction using random forest classifier with different wrapper methods. In 2022 International Conference on Edge Computing and Applications, Tamilnadu, India, pp. 1705-1710. <https://doi.org/10.1109/ICECAA55415.2022.9936172>
- [12] Zhu, T., Li, K., Herrero, P., Georgiou, P. (2021). Deep Learning for diabetes: A systematic review. *IEEE Journal of Biomedical and Health Informatics*, 25(7): 2744-2757. <https://doi.org/10.1109/JBHI.2020.3040225>
- [13] Kaur, H., Kumari, V. (2022). Predictive modelling and analytics for diabetes using a machine learning approach. *Applied Computing and Informatics*, 18(1/2): 90-100. <https://doi.org/10.1016/j.aci.2018.12.004>
- [14] Chien, T.Y., Ting, H.W., Chen, C.F., Yang, C.Z., Chen, C.Y. (2022). A clinical decision support system for diabetes patients with deep learning: Experience of a Taiwan medical center. *International Journal of Medical Sciences*, 19(6): 1049-1055. <https://doi.org/10.7150/ijms.71341>
- [15] Fazakis, N., Kocsis, O., Dritsas, E., Alexiou, S., Fakotakis, N., Moustakas, K. (2021). Machine learning tools for long-term type 2 diabetes risk prediction. *IEEE Access*, 9: 103737-103757. <https://doi.org/10.1109/ACCESS.2021.3098691>
- [16] Mehta, R., Vala, B., Patel, A. (2022). A survey on diabetes prediction using supervised learning. In Proceedings of the 2nd International Conference on Artificial Intelligence and Smart Energy (ICAIS 2022), Coimbatore, India, pp. 302-307. <https://doi.org/10.1109/ICAIS53314.2022.9743006>
- [17] Sari, F.A.O., Alrammahi, A.A.H., Hameed, A.S., Alrikabi, H.M.B., Abdul-Razaq, A.A., Nasser, H.K., AL-Rifaie, M.F. (2022). Networks cyber security model by using machine learning techniques. *International Journal of Intelligent Systems and Applications in Engineering*, 10(3s): 257-263.
- [18] Alrifai, M.F., Ahmed, Z.H., Hameed, A.S., Mutar, M.L. (2021). Using machine learning technologies to classify and predict heart disease. *International Journal of Advanced Computer Science and Applications*, 12(3): 123-127. <https://doi.org/10.14569/IJACSA.2021.0120315>
- [19] Rathod, S.R., Phadke, L., Chaskar, U.M., Patil, C.Y. (2021). Machine learning techniques for predicting Type 2 diabetes mellitus risk using heart rate variability features. In 2021 12th International Conference on Computing Communication and Networking Technologies, Kharagpur, India, pp. 1-6. <https://doi.org/10.1109/ICCCNT51525.2021.9579746>
- [20] Qin, Y.F., Wu, J.L., Xiao, W., Wang, K., Huang, A.B., Liu, B., Yu, J.X., Li, C., Yu, F.Y., Ren, Z.B. (2022). Machine learning models for data-driven prediction of diabetes by lifestyle type. *International Journal of Environmental Research and Public Health*, 19(22): 15027. <https://doi.org/10.3390/ijerph192215027>