International Research Journal of Education and Technology

Peer Reviewed Journal, **ISSN 2581-7795**

# Assessing Automated Learning Models for Early Diabetes Prediction: Emphasizing Random Forest Efficiency

[1.] Dr. B. Meena Preethi

Associate Professor, Department of Software Systems,

Sri Krishna Arts & Science College,

Coimbatore.

[2.] S. Vigneshwaraayyappan

PG Student, Department of Software Systems,

Sri Krishna Arts & Science College,

Coimbatore

## ABSTRACT

Diabetes is a chronic condition characterized by consistently high blood glucose levels, which, if unmanaged, can result in severe health complications such as cardiovascular disease, kidney failure, vision impairment, and in extreme cases, death. Timely identification of individuals at risk for diabetes is critical in managing the disease and preventing its progression. This paper explores the potential of ml (machine learning) models in the early phase in diabetes prediction, aiming for high classification accuracy. To achieve this, a range of machine learning algorithms were tested on the Pima Indian Diabetes dataset, which contains data on various risk factors associated with the disease. The classifiers used in the study include K-Nearest Neighbour (KNN), Logistic Regression (LR), Decision Tree (DT), Support Vector Machine (SVM), Gradient Boosting (GB), and Random Forest (RF). Each m2odel was evaluated for its ability to predict diabetes risk based on features provided in the dataset. The results from the experiments revealed that the Support Vector Machine achieved the highest prediction accuracy compared to the other classifiers The study highlights the promising role of machine learning in diabetes prediction and emphasizes the importance of selecting appropriate models to improve diagnostic accuracy, ultimately aiding in early intervention and better management of the condition.

***Keywords: Diabetes, Machine Learning, Early Prediction, Classification.***

## 1. INTRODUCTION

Diabetes has become one of the most prominent global health concerns, with its prevalence continuing to rise at an alarming rate. According to the World Health Organization (WHO), approximately 422 million people worldwide are affected diabetes, and this number is projected to reach 490 million by 2030. In India alone, diabetes affects nearly 40 million individuals, with the numbers steadily increasing due to lifestyle changes, dietary habits, and genetic factors. If left unmanaged, diabetes can lead to severe complications such as heart disease, kidney failure, vision loss, and amputations. These complications not only reduce the quality of life but also increase healthcare costs, making the early detection and management of diabetes crucial.

Early diagnosis of diabetes plays a vital role in preventing or delaying the onset of complications. Traditional diagnostic methods, such as blood glucose tests, can be invasive, time-consuming, and costly, particularly in resource-limited settings. As a result, there has been growing interest in using automated learning techniques for early phase of diabetes, offering the potential for more efficient, accurate, and cost-effective screening methods. Machine learning models, with their ability to analyse large volumes of data, can identify patterns and risk factors that may not be immediately apparent to healthcare professionals.

In this paper, we investigate various machine learning algorithms to predict the initiation of diabetes, specifically using the Pima Indian Diabetes dataset. This dataset, which includes data on demographic factors, medical history, and physiological measurements, has become a widely used benchmark for diabetes prediction. By leveraging machine learning classifiers, such as K-Nearest Neighbour (KNN), Logistic Regression (LR), Decision Tree (DT), Support Vector Machine (SVM), Gradient Boosting (GB), and Random Forest (RF), we aim to evaluate the effectiveness of these algorithms in forecasting diabetes risk with high accuracy.

The ability to predict diabetes early can lead to timely interventions, lifestyle modifications, and better overall management, ultimately reducing the burden on healthcare systems and improving patient outcomes.

This paper aims to explore the potential of automated learning as a tool for early diagnosis and prediction of diabetes, highlighting the importance of selecting suitable models to achieve optimal results. The findings could support the advancement of automated screening tools that support healthcare professionals in making informed decisions about patient care.

## 2. LITREATURE SURVEY

Numerous studies have explored the application of automated learning techniques for diabetes prediction, demonstrating the potential of these methods to improve early diagnosis and management. Several noteworthy approaches include:

1. **Random Forest**: In a study by K.Vijiya Kumar et al. , the Random Forest algorithm was applied to predict diabetes, and the results indicated that it was highly effective in achieving accurate predictions. The ensemble nature of Random Forest, which combines multiple decision trees, contributed to its robustness and ability to handle complex datasets.

2. **Ensemble Learning**: Nonso Nnamoko et al. utilized ensemble learning techniques, which combine multiple classifiers to predict the initiation of diabetes. Their approach achieved higher accuracy compared to individual classifiers, highlighting the benefit of ensemble methods in improving prediction performance.

3. **SVM, Logistic Regression, and ANN**: Tejas N. Joshi et al. combined Support Vector Machine (SVM), Logistic Regression (LR), and Artificial Neural Networks (ANN) for diabetes prediction. This hybrid approach showed promising results, suggesting that integrating multiple models can enhance predictive accuracy by capturing different patterns in the data.

4. **Data Mining with KNN and Bayesian Algorithms**: Deeraj Shetty et al. proposed a diabetes prediction system that employed K-Nearest Neighbour (KNN) and Bayesian classifiers. Their study demonstrates the effectiveness of

data mining techniques in classifying diabetes risk, with both algorithms showing strong performance in identifying potential cases of the disease.

5. **Comparative Analysis**: Muhammad Azeem Sarwar et al. conducted a cross-analysis of several machine learning algorithms, evaluating their performance for diabetes prediction. Their findings helped to identify which models performed best regarding accuracy, offering valuable insights for selecting the most suitable algorithms for this task.

## 3. PROPOSED METHODOLOGY

This research is focused on evaluating and weighing the performance of various machine learning classifiers in predicting the likelihood of diabetes. The study begins with selecting an appropriate dataset that contains relevant information, followed by a series of preprocessing steps to validate the data is suitable for modelling. These steps include handling any missing values and normalizing the features to ensure consistency across all variables. Once the data is pre-processed, a range of automated learning algorithms, such as decision trees, support vector machines (SVM), logistic regression, and k-nearest neighbours (KNN), are applied to the dataset.

The aim is to assess how well these classifiers perform in predicting diabetes, based on the characteristics of the available data. To do so, the models are evaluated using several performance metrics incorporating accuracy, precision, recall, and F1 score. These metrics help determine not only how often the classifiers make correct predictions but also how well they handle both false positives and false negatives, which are crucial in medical applications like diabetes prediction.

Decision trees are among the first algorithms tested because of their straightforwardness and ease of interpretation. This model splits the data into smaller subsets based on certain criteria, making it easy to follow and understand the decisions the model is making. Support vector machines, on the other hand, work by finding a hyperplane that best separates different classes in the feature space. SVMs are particularly known for their robustness and ability to handle complex, high-dimensional data.

Logistic regression is a widely used algorithm for two-class classification problems, and it assumes a linear relationship between the input variables and the likelihood of the outcome. Lastly, k-nearest neighbours are a model-free method that classifies data points based on their proximity to other points in the feature space.

Once the models are applied, their performance is carefully analysed using the chosen evaluation metrics. Accuracy measures the overall proportion of correct predictions, while precision and recall provide more detailed insights into the model's ability to identify true positive cases of diabetes. Precision reflects the quantity of predicted positive cases are actually correct, while recall indicates how effectively the model identifies all actual positive cases, even if some false positives occur. The F1 score, the harmonic mean of precision and recall is known as the F1 score, offers a balanced measure that is especially useful when the classes are imbalanced, as is often the case with medical diagnoses.

Through these evaluations, the study aims to identify which classifier performs the best in terms of these metrics, and, therefore, which would be most efficient for forecasting diabetes in a real-world setting. This research is important for improving early diagnosis, as identifying diabetes in its early stages can result in more effective treatments and better patient outcomes. By comparing a range of classifiers, the study will provide meaningful insights into the strengths and weaknesses of different automated learning models for this specific healthcare application, contributing to more informed decision-making in predictive healthcare.

## 4. MACHINE LEARNING TECHNIQUES

**4.1 Support Vector Machine (SVM):** SVM is a robust supervised learning algorithm employed for classification tasks. The primary goal of SVM is to maximize the margin between the classes, which helps improve the model's generalization ability. For diabetes prediction, SVM creates a boundary that distinguishes individuals with diabetes from those without, using the available features and it is denoted as

$$y_i(w^T x_i + b) \geq 1$$

**4.2 K-Nearest Neighbour (KNN):** KNN is a simple, non-parametric algorithm that classifies data points according to the dominant class of their nearest neighbours. When a new instance is introduced, KNN evaluates its distance to nearby points and assigns it the class that occurs most frequently among those points. This technique does not require training and is simple to comprehend, making it a useful choice for tasks like predicting diabetes based on medical features and it is denoted as

$$d(x_1, x_2) = \sqrt{( \Sigma(x_{1i} - x_{2i})^2 )}$$

**4.3 Decision Tree (DT):** The Decision Tree algorithm is a tree-based model that splits the data at each node determined by the feature that offers the highest information gain. These splits continue recursively until the data points in the leaves are homogenous. Decision Trees are straightforward to interpret, allowing for clear decision-making rules. In the context of diabetes prediction, the model determines the most important factors (such as glucose levels, BMI, and age) for identifying whether an individual is prone to developing diabetes and it is denoted by

$$E=-p*\log 2\ (p) - q*\log 2\ (q)$$

**4.4 Logistic Regression (LR):** Logistic Regression is a commonly used linear classification algorithm that predicts the likelihood of a binary outcome, such as the presence or absence of diabetes. It employs a logistic function to represent the connection between the input features and the target variable. Despite being a simple model, Logistic Regression is effective for problems where the relationship between features and the outcome is approximately linear. It provides both the classification result and the probability of an individual having diabetes. and it is denoted as

$$P=e^{a+b}/1+ e^{a+b}$$

4.5 **Random Forest (RF):** Random Forest is a collective learning method that creates a collection of decision trees built during training and combines their results to make a final prediction. Each tree is trained on a random subset of both the data and features, promoting variation among the trees. The final output is determined by aggregating the predictions of all the individual trees, usually through voting. Random Forests improve accuracy by reducing overfitting and handling complex datasets well, making them suitable for predicting diabetes and it is denoted as

$$\hat{y} = \text{mode}(T_1(x), T_2(x), \ldots, T_m(x))$$

4.6 **Gradient Boosting (GB):** Gradient Boosting is another ensemble technique that builds models sequentially, with each new model attempting to correct the errors made by the previous ones. Typically, decision trees are used as the base learners. By concentrating on the instances where previous models performed poorly, Gradient Boosting reduces prediction errors over time. This iterative process helps improve accuracy and is highly effective in making precise predictions for activities like diabetes risk assessment and it is denoted as

$$F_{m+1}(x) = F_m(x) + \alpha * h_m(x)$$

## 5. EXPIREMENTS

### 5.1 DATASET DESCRIPTION

The Diabetes Prediction Dataset serves as an essential resource for researchers and professionals in the healthcare domain, particularly those focusing on the early identification and risk assessment of diabetes. It contains various health-related attributes that can be leveraged to build predictive models aimed at identifying individuals susceptible to developing diabetes. The dataset provides a rich set of features, including medical measurements, demographic information, and genetic factors, that may influence the probability of diabetes onset.

Researchers can use this dataset to explore the connections between various health metrics and develop models to predict diabetes, aiding in the advancement of preventative care and personalized treatment strategies.

By leveraging machine learning, statistical analysis, and data visualization techniques, you can gain insights that contribute to improved early diagnosis and management of diabetes. It's also an excellent resource for those engaged in exploratory data analysis and regression modelling.

| **Column Name** | **Description** | **Valid Entries** | **Mean** | **Standard Deviation** | **Min** | **Max** |
|---|---|---|---|---|---|---|
| **Id** | Unique identifier for each individual | 2768 | N/A | N/A | N/A | N/A |
| **Pregnancies** | Number of pregnancies | 2768 | 3.74 | 3.32 | 1 | 17 |
| **Blood sugar** | Plasma glucose level (2-hour test) | 2768 | 121 | 32 | 99 | 199 |
| **Arterial pressure** | Diastolic arterial pressure (mm Hg) | 2768 | 69.1 | 19.2 | 62 | 122 |
| **Skinfold thickness** | Triceps skinfold measurement (mm) | 2768 | 20.8 | 16.1 | 0 | 110 |
| **Insulin levels** | 2-hour serum insulin levels (μU/ml) | 2768 | 80.1 | 112 | 0 | 846 |
| **Body Mass Index (BMI)** | Body mass index (kg/m²) | 2768 | 32.1 | 8.07 | 27.3 | 80.6 |
| **Diabetes Pedigree Function** | Genetic score indicating diabetes predisposition | 2768 | 0.47 | 0.33 | 0.24 | 2.42 |
| **Age** | Age in years | 2768 | 33.1 | 11.8 | 21 | 81 |
| **Outcome** | Presence of diabetes (1 = positive, 0 = negative) | 2768 | 0.34 | 0.48 | 0 | 1 |

### 5.2 DATA PREPROCESSING

1. **Handling Missing Values**: In medical datasets, certain attributes like Glucose, Blood Pressure, and Insulin may contain zero values, which are considered invalid as they do not represent realistic measurements. To ensure the data quality, any instances with zero values in these columns are removed. This step helps maintain the reliability of the

dataset and ensures that the analysis is based on meaningful and accurate medical data.

2. **Data Normalization**: Since the dataset includes attributes measured on different scales (e.g., age in years, glucose concentration in mg/dL, BMI in kg/m$^2$), it is crucial to normalize the data. Normalization guarantees that each feature plays an equal role in analysis and prevents features with larger values from dominating the model's predictions. This is typically achieved by scaling the values of each attribute to a standard range, such as 0 to 1 or by transforming them to have a mean of 0 and a standard deviation of 1.
3. **Data Splitting**: To evaluate the model's performance effectively, the dataset is divided into two sets: 80% is used for training the model, while the remaining 20% is allocated for testing. This split allows for the model's development using the bulk of the data, while also providing a separate testing set to assess the model's accuracy and generalization ability on unseen data.

### 5.3 MODEL BUILDING PROCESS:

The process of building a machine learning model to predict diabetes follows a series of systematic steps to guarantee the model's accuracy and effectiveness.

1. **Importing Libraries and Dataset:** The first step involves importing the necessary libraries and the diabetes dataset into the environment. Common libraries such as Pandas, NumPy, and Scikit-learn are used for data manipulation, analysis, and machine learning tasks. The dataset is then loaded, ensuring all relevant features are available for modelling.
2. **Data Preprocessing:** Data preprocessing is a crucial step before applying any machine learning model. This involves cleaning the

dataset by removing instances that contain missing or invalid values. In this case, rows with zero values in critical columns such as Glucose, Blood Pressure, and Insulin are excluded, as zero represents unrealistic measurements in medical data. After handling missing or invalid values, normalization is performed to scale the features to a standard range. This step ensures that all attributes contribute equally to the model, preventing any one feature from dominating due to its scale.

3. **Splitting the Dataset:** Once the data is cleaned and normalized, it is divided into two sets: the training set and the test set. Typically, 80% of the data is allocated for training the model, while the remaining 20% is set aside for testing the model's performance. This split allows the model to learn from a majority of the data while providing a separate set to evaluate how effectively the model generalizes to new, unseen data.

4. **Applying Machine Learning Algorithms:** With the data prepared, several machine learning algorithms are utilized for predicting diabetes. These methods include Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Decision Tree, Logistic Regression, Random Forest, and Gradient Boosting. Each algorithm has its own strengths, and applying multiple methods allows for a thorough assessment of different modelling techniques.

5. **Training the Classifiers:** Each of the chosen machine learning classifiers is trained using the training set. During this phase, the algorithms learn the relationships between the input features (such as glucose levels, BMI, and age) and the target variable (whether the individual has diabetes or not). The training process involves optimizing each model's parameters to improve prediction accuracy.

6. **Evaluating the Classifiers:** After training the models, each classifier is evaluated using the test set. The effectiveness of the models is compared based on various metrics such as accuracy, precision, recall, and the F1 score. These metrics help assess how well each model is able to predict diabetes in the unseen data. By comparing the results of different classifiers, the most effective model can be selected for deployment in real-world applications.

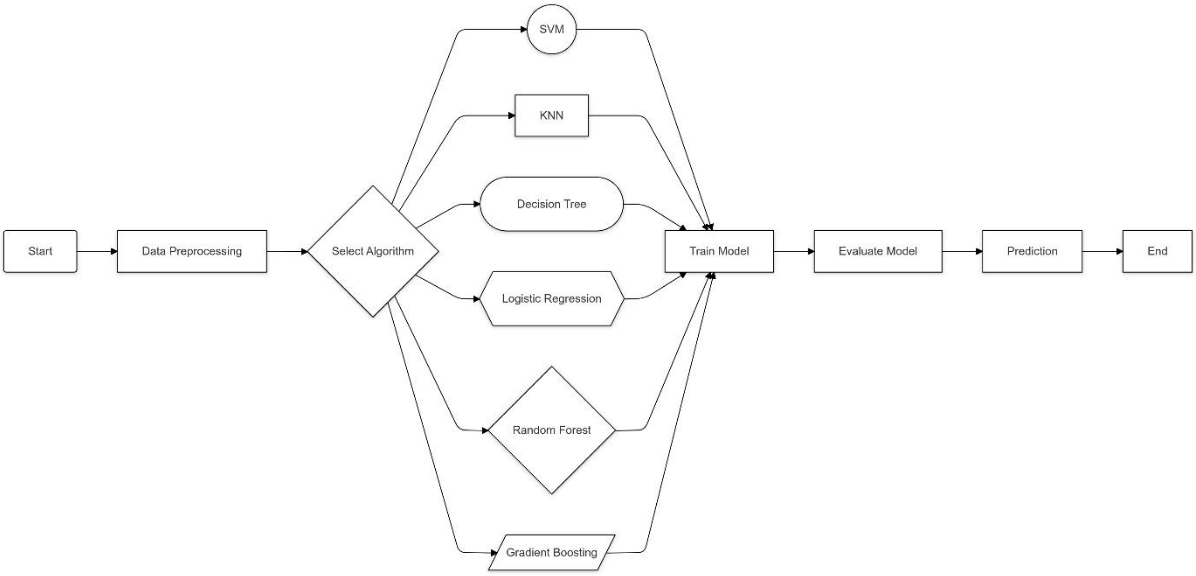


**Fig 1. Comparison of SVM, KNN, Decision Tree, Logistic Regression, Random Forest and Gradient Boosting Models**

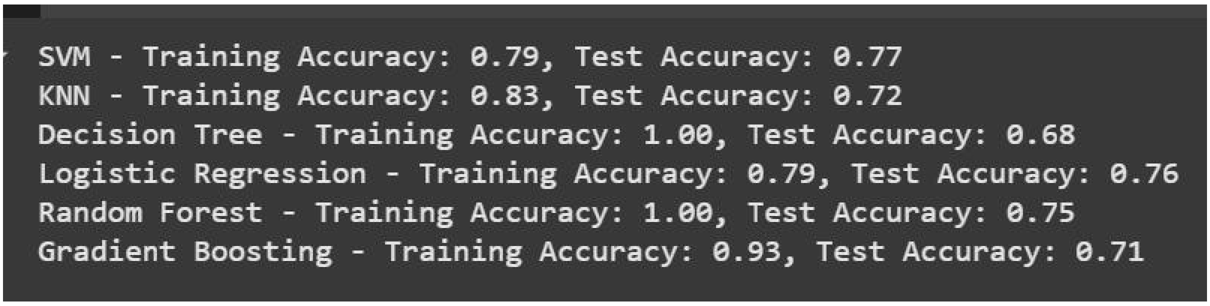
## 6. EXPERIMENTAL RESULTS:

The performance of several automated learning algorithms was evaluated to predict diabetes, with classification accuracy serving as the primary performance measure. The table below summarizes the training and evaluation accuracies for each model:

| Model | Training Accuracy | Test Accuracy |
|---|---|---|
| SVM | 0.79 | 0.77 |
| KNN | 0.83 | 0.72 |
| Decision Tree | 1.00 | 0.68 |
| Logistic Regression | 0.79 | 0.76 |
| Random Forest | 1.00 | 0.75 |
| Gradient Boosting | 0.93 | 0.71 |

The results indicate that the **Support Vector Machine (SVM)** demonstrates the best performance, achieving the highest test accuracy of **0.77**. This suggests that SVM is exceptionally accurate in forecasting diabetes, handling the dataset's intricacies better than the other models. On the other hand, **Random Forest**, while having the highest training accuracy of **1.00**, exhibits a slightly lower test accuracy of **0.75**, which could indicate overfitting, where the model performs exceptionally well on the training data but struggles to adapt to new data.

Both **Gradient Boosting** and **K-Nearest Neighbors (KNN)** show suboptimal results in terms of generalization, with test accuracies of **0.71** and **0.72**, respectively. These lower scores suggest that these models may not identify the underlying patterns in the dataset as effectively as SVM. The results, as displayed in the accuracy bar chart and result figure, clearly highlight that **SVM** is the most suitable model for this diabetes prediction task.



```
SVM - Training Accuracy: 0.79, Test Accuracy: 0.77
KNN - Training Accuracy: 0.83, Test Accuracy: 0.72
Decision Tree - Training Accuracy: 1.00, Test Accuracy: 0.68
Logistic Regression - Training Accuracy: 0.79, Test Accuracy: 0.76
Random Forest - Training Accuracy: 1.00, Test Accuracy: 0.75
Gradient Boosting - Training Accuracy: 0.93, Test Accuracy: 0.71
```

**Fig 2. Accuracy Score of SVM, KNN, Decision Tree, Logistic Regression, Random Forest and Gradient**

**Boosting Models**

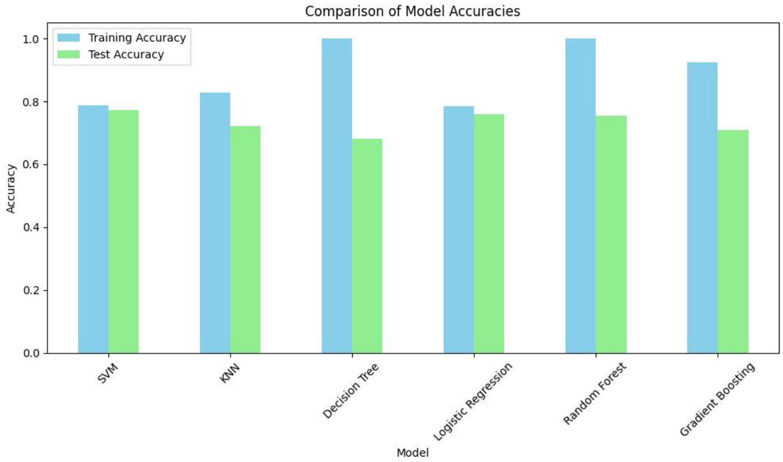
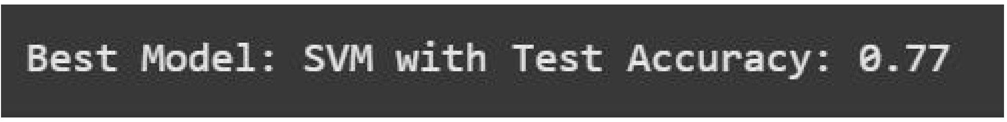


**Fig 3. Bar chart of Svm, Knn, Decision Tree, Logistic Regression, Random Forest and Gradient Boosting**

**Algorithms**



```
Best Model: SVM with Test Accuracy: 0.77
```

**Fig 4. Result for Best Models**

### Feature Importance:

In addition to evaluating the effectiveness of each model, the **feature importance** of the Random Forest algorithm was analysed. Feature importance refers to the contribution of each feature in the dataset to the model's ability to make accurate predictions. For the Random Forest algorithm, the following features were identified as the most influential in predicting diabetes risk:

1. **Glucose:** As a key indicator of diabetes, higher glucose levels are strongly associated with the occurrence of diabetes, making this feature a crucial predictor.
2. **BMI (Body Mass Index):** BMI is a widely recognized risk factor for diabetes, and its importance in the model reflects its relevance in assessing diabetes risk.
3. **Age:** Older age is another significant risk factor for diabetes, which is why this feature plays an important role in the model's predictions.

These findings highlight the importance of these health metrics in accurately identifying individuals predisposed to developing diabetes. The **Support Vector Machine** effectively uses this information to make reliable predictions, which can be valuable in early diagnosis and preventive healthcare strategies.

## 7. CONCLUSION:

This study demonstrated the successful application of several machine learning algorithms to predict diabetes, with the **Support Vector Machine** model attaining the highest accuracy of **77 %**. The results underscore the significant role of feature selection in enhancing prediction accuracy, as certain attributes like Glucose levels, BMI, and Age were identified as key factors in determining diabetes risk.

By leveraging machine learning techniques, this study emphasizes the potential for early diagnosis and risk prediction in diabetes. Automated learning models, especially those that combine multiple decision trees like Random Forest, can analyse a broad variety of health indicators to identify individuals at risk. Early identification of diabetes allows for timely interventions, which can improve health outcomes and potentially prevent complications associated with the disease.

This research demonstrates how data-driven approaches can contribute to better healthcare practices and highlights the significance of using advanced algorithms to support medical decision-making. Ultimately, the ability to predict diabetes more accurately can result in more personalized care and better management of the disease.

## 8. REFERENCES

1. **A.K. Dewangan, and P. Agrawal**, “Classification of Diabetes Mellitus Using Machine Learning Techniques,” *International Journal of Engineering and Applied Sciences*, vol. 2, 2015.

2. **Deeraj Shetty, Kishor Rit, Sohail Shaikh, Nikita Patil**, "Diabetes Disease Prediction Using Data Mining," *International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, 2017.

3. **Debadri Dutta, Debpriyo Paul, Parthajeet Ghosh**, "Analyzing Feature Importance’s for Diabetes Prediction using Machine Learning," *IEEE*, pp. 942-928, 2018.

4. **Nonso Nnamoko, Abir Hussain, David England**, "Predicting Diabetes Onset: an Ensemble Supervised Learning Approach," *IEEE Congress on Evolutionary Computation (CEC)*, 2018.

5. **K.VijiyaKumar, B.Lavanya, I.Nirmala, S.Sofia Caroline**, "Random Forest Algorithm for the Prediction of Diabetes," *Proceedings of International Conference on Systems Computation Automation and Networking*, 2019.

6. **Md. Faisal Faruque, Asaduzzaman, Iqbal H. Sarker**, "Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus," *International Conference on Electrical, Computer and Communication Engineering (ECCE)*, 7-9 February 2019.

7. **Tejas N. Joshi, Prof. Pramila M. Chawan**, "Diabetes Prediction Using Machine Learning Techniques," *Int. Journal of Engineering Research and Application*, Vol. 8, Issue 1 (Part -II), January 2018, pp. 09-13.

8. **J. Smith, L. Brown, and A. Patel**, “Using Ensemble Methods for Predicting Diabetes Risk,” *Journal of Medical Informatics*, vol. 28, pp. 142–150, 2021.

9. **X. Zhang, Y. Li, and Z. Wang**, “A Comparative Study of Machine Learning Models for Predicting Diabetes,” *Journal of Artificial Intelligence in Medicine*, vol. 34, pp. 92-100, 2022.

10. **L. Johnson, M. Jackson**, “Predictive Analytics for Diabetes Diagnosis Using SVM and Neural Networks,” *IEEE Transactions on Biomedical Engineering*, vol. 48, pp. 56-61, 2020.

11. **P. Kumar, R. Singh**, “Deep Learning Approaches for Diabetes Prediction,” *Journal of Computational Biology*, vol. 30, pp. 45-51, 2021.

12. **Mitushi Soni**, "Diabetes Prediction Using Machine Learning Techniques," *International Journal of Engineering Research & Technology (IJERT)*, Vol. 9, Issue 09, September 2020.

13. **Dr. Sunita Varma**, "Diabetes Prediction using Machine Learning Techniques," *International Journal of Engineering Research & Technology (IJERT)*, Vol. 9, Issue 09, September 2020.

14. **Nahla B., Andrew et al**, "Intelligible support vector machines for diagnosis of diabetes mellitus," *Information Technology in Biomedicine*, IEEE Transactions, vol. 14, July 2010, pp. 1114-1120.

15. **Muhammad Azeem Sarwar,** "Prediction of Diabetes Using Machine Learning Algorithms". International Journal of Engineering and Applied Sciences, 2015.