Cureus Journal of Computer Science
https://doi.org/10.7759/s44389-025-00058-8

Original Article

Enhancing Early Diabetes Detection Through Machine Learning: Analyzing Data for Accurate Risk Prediction

Amita Shukla [1], Garima Jain [2], Sovers Singh Bisht [3], Anshika Panwar [4], Meet Thakur [2, ✉], Ankush Jain [5]

1. *Computer Science and Engineering (Artificial Intelligence), IIMT Group of Colleges, Greater Noida, IND*
2. *Computer Science and Business Systems, Noida Institute of Engineering and Technology, Greater Noida, IND*
3. *Data Science, Noida Institute of Engineering and Technology, Greater Noida, IND*
4. *Electronics & Communications Engineering, Bharati Vidyapeeth's College of Engineering, New Delhi, IND*
5. *Computer Science and Engineering, Netaji Subhas Institute of Technology, Delhi, IND*

Received: August 30, 2025 | Review began: February 04, 2026 | Review ended: April 13, 2026 | Published: April 28, 2026

Abstract

Diabetes is a chronic condition that disrupts the body's ability to regulate blood sugar levels. If left untreated, it can lead to serious health complications. Early identification and timely intervention are important to prevent these adverse outcomes. This research paper examines the use of machine learning algorithms to analyze large datasets, including demographic information, medical history, and laboratory test results, to identify diabetes-related patterns and risk factors. The focus is on supervised learning, where the algorithm is trained on labeled data with diabetes to differentiate between individuals. Once trained, these models classify new individuals based on their characteristics and estimate their risk of developing the condition. The study applies three separate machine learning techniques to predict the onset of diabetes with high accuracy. By integrating diverse data sources and using cross-validation, the model achieved strong predictive performance. The main objective was to detect diabetes early and evaluate the effectiveness of machine learning in this context. Results suggest that these techniques significantly improve the accuracy and reliability of diabetes prediction, enabling proactive care and potentially better health outcomes for at-risk individuals.

Categories: AI applications, AI/ML-based decision support systems, Health Informatics

Keywords: diabetes, early detection, machine learning algorithms, classification, prediction

Introduction

Diabetes is a non-communicable disease that poses a significant threat to global health [1]. It arises from either insufficient insulin production by the pancreas or the body's inefficient use of insulin. According to the World Health Organization (WHO), the number of individuals with diabetes is projected to reach 552 million by 2030, impacting one in every ten adults. This alarming trend underscores the urgent need for effective prevention and management strategies [2]. Predicting diabetes using machine learning algorithms involves creating models that can accurately determine if an individual has diabetes based on personal characteristics [3]. These models are trained on data collected from both diabetic and non-diabetic patients and can then make predictions on new, unseen data. Key features in the dataset may include blood pressure, age, gender, body mass index (BMI), and family history of the disease. Different machine learning methods like logistic regression, decision trees, random forests, and support vector machines (SVM) can be used for this problem [3]. The primary purpose of employing machine learning for diabetes prediction is to increase early diagnosis and prevention [4]. By predicting the possibility of diabetes, doctors can first interfere with strategies to delay or prevent its onset. Early diagnosis and prevention can improve health results by reducing the risk of complications such as heart

disease, kidney disease, blindness, and dissection. It also reduces the economic and social burden associated with the losses related to diabetes and the loss of productivity [5]. Machine learning algorithms uncover patterns and relationships in patient data that may not be immediately apparent to humans, enabling effective self-management and intervention strategies. Diabetes detection systems have many applications [6]. This allows individuals to monitor their blood sugar levels at home and help healthcare professionals to provide more effective treatment. Early diagnosis enables patients to make necessary dietary and lifestyle changes sooner. This system also offers statistical insights on different supervised learning algorithms like SVM, decision trees, and random forests [7]. In conclusion, machine learning techniques can estimate an individual's risk of developing diabetes, thereby improving early detection, prevention, and control strategies, and ultimately enhancing the health of those at risk for or affected by diabetes.

Diabetes detection using machine learning algorithms has been extensively studied. Choosing the right model is important because it can affect the accuracy and validity of predicting diabetes. Multivariate models are better for certain data types and forecasting. For example, the decision tree model will be better at predicting blood glucose based on the number of binary determinants, while the SVM will produce better results with larger estimates [8]. Model has a significant impact on diabetes prognosis by analyzing quantitative data to identify patterns and reveal the truth. This is especially important given the many risk factors that lead to diabetes, such as age, family history, obesity, and physical inactivity [9]. Machine learning has many ways to predict diabetes. Supervised learning involves training a model. Unsupervised learning uses unlabeled data to find patterns or clusters, whereas supervised learning uses labeled data with known values. Other methods include cluster learning, which combines multiple models to increase accuracy, and deep learning, which uses neural networks to process and analyze fixed data [10]. When using machine learning for diabetes prediction, several challenges and potential issues arise, including data quality, workload, class disparities, privacy and security concerns, information availability, information inconsistencies, diabetes pressure, and inability to interpret gender. Methods such as decision trees, random forests, SVM, and deep learning have shown potential for predicting diabetes. By using various types of data and different machine learning algorithms, the accuracy of these predictions can be improved. However, more research is still needed to find out the most effective way to use machine learning for predicting the risk of diabetes [11].

Materials And Methods

This research paper focuses on enhancing the precision of diabetes testing methods using machine learning techniques. Experiments were conducted with various classification and clustering algorithms to achieve this objective [1,2]. The following section provides an overview of the methodology adopted throughout the study.

The process begins with importing the required libraries and the diabetes dataset obtained from the UCI repository [9,12]. Data preprocessing is prioritized to handle missing and inconsistent values, which is a crucial step in healthcare datasets [1,4]. The dataset is then divided into training data (80%) and testing data (20%) to ensure unbiased performance evaluation [10]. Classification models are built using the training dataset for each selected machine learning algorithm, and their performance is evaluated using standard metrics such as accuracy, precision, and recall [3,7]. A comparative analysis is then carried out to determine the most efficient algorithm based on multiple evaluation criteria [8].

How to cite this article:
Shukla A, Jain G, Singh Bisht S, et al. (April 28, 2026) Enhancing Early Diabetes Detection Through Machine Learning: Analyzing Data for Accurate Risk Prediction. Cureus J Comput Sci 3 : es44389-025-00058-8. DOI https://doi.org/10.7759/s44389-025-00058-8

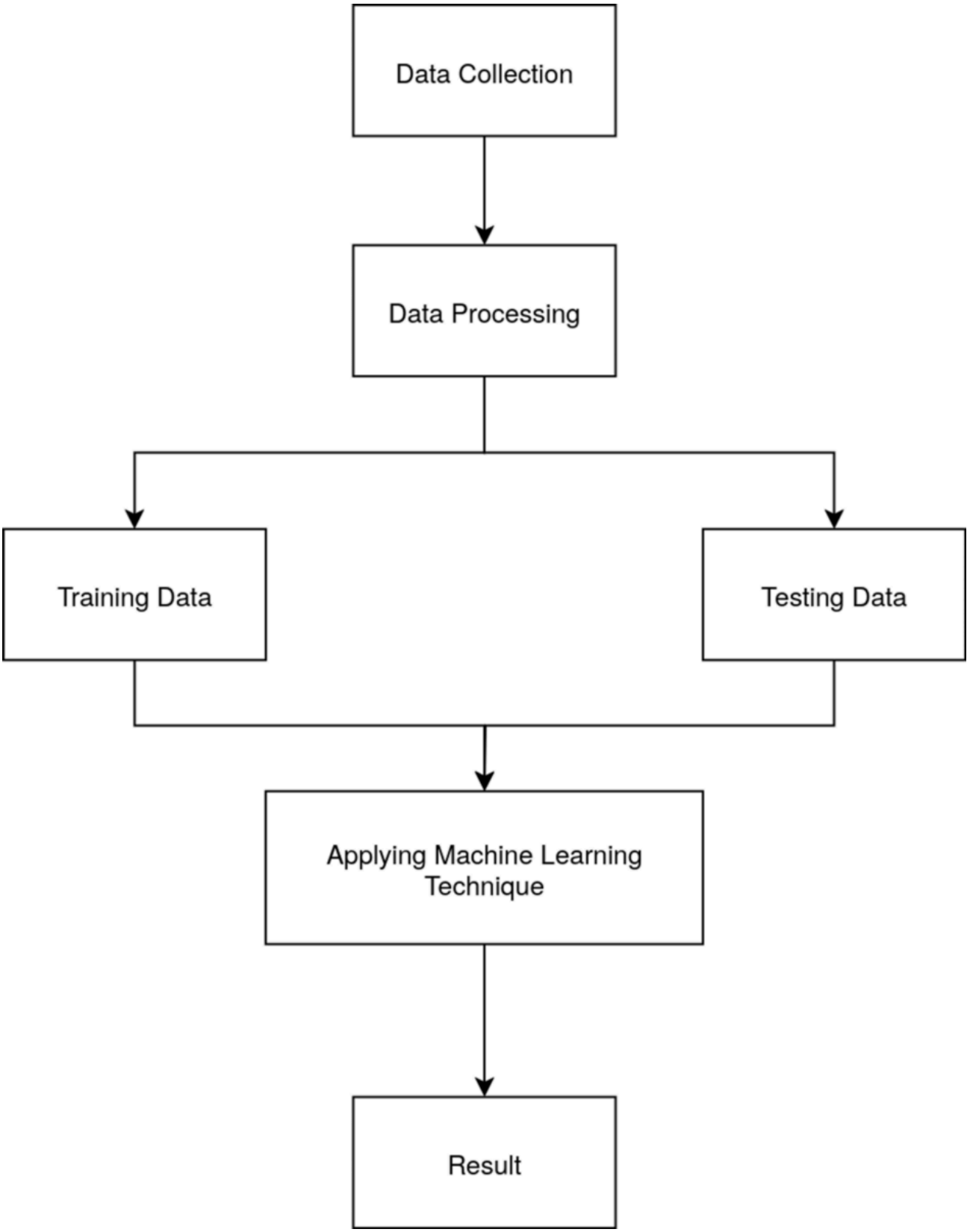

FIGURE 1: Design Approach

How to cite this article:
Shukla A, Jain G, Singh Bisht S, et al. (April 28, 2026) Enhancing Early Diabetes Detection Through Machine Learning: Analyzing Data for Accurate Risk Prediction. Cureus J Comput Sci 3 : es44389-025-00058-8. DOI https://doi.org/10.7759/s44389-025-00058-8

Figure *1* presents the overall design approach adopted in this study. The methodology follows a structured six-phase pipeline. In the first phase, Raw Data Collection, the Pima Indian Diabetes Dataset is obtained from the UCI Machine Learning Repository, comprising medical diagnostic measurements for female patients including glucose levels, blood pressure, BMI, and age. The second phase, Data Preprocessing, involves removal of missing and zero values for medically implausible attributes, followed by normalization and feature selection to retain the five most clinically significant variables: Glucose, Blood Pressure, Skin Thickness, BMI, and Age. In the third phase, the preprocessed dataset undergoes Data Splitting, where 80% of records are allocated for training and the remaining 20% are reserved for testing to ensure unbiased performance evaluation. The fourth phase covers Model Training, in which three supervised machine learning algorithms - SVM, Decision Tree, and Random Forest - are independently trained on the training subset. The fifth phase involves Model Evaluation, where each trained model is assessed on the test subset using standard performance metrics including Accuracy, Precision, Recall, and F1-Score. Finally, the sixth phase represents Prediction Output and Deployment, where the best-performing model is integrated into a decision-support pipeline capable of real-time diabetes risk prediction.

Dataset description

The dataset used in this study is obtained from the UCI Machine Learning Repository and is commonly known as the Pima Indian Diabetes Dataset [9,11]. It consists of medical diagnostic measurements for female patients of Pima Indian heritage. The ninth attribute represents the class label, where 0 indicates non-diabetic and 1 indicates diabetic. Dataset description is presented in Table *1*.

S. No.	Attributes
1	Pregnancy
2	Glucose
3	Blood Pressure
4	Skin Thickness
5	Insulin
6	Body Mass Index
7	Diabetes Pedigree Function
8	Age

TABLE 1: Dataset Description

Data preprocessing

Data preprocessing is a critical phase, particularly in healthcare applications, as raw data often contains missing values, noise, and inconsistencies [1,6]. To improve the effectiveness of machine learning models, preprocessing techniques are applied to clean and prepare the dataset.

How to cite this article:
Shukla A, Jain G, Singh Bisht S, et al. (April 28, 2026) Enhancing Early Diabetes Detection Through Machine Learning: Analyzing Data for Accurate Risk Prediction. Cureus J Comput Sci 3 : es44389-025-00058-8. DOI https://doi.org/10.7759/s44389-025-00058-8

Missing Values Removal: Attributes containing zero values for features such as glucose level, blood pressure, skin thickness, and BMI are treated as missing, as such values are medically implausible. These records are removed to improve model reliability [4,9]. Feature selection is also applied to reduce dimensionality and improve computational efficiency [6].

Data Splitting: After cleaning and normalization, the dataset is divided into training and testing subsets. The training data is used to learn model parameters, while the test data is reserved for evaluating model generalization [10].

S. No.	Attributes
1	Glucose
2	Blood Pressure
3	Skin Thickness
4	Body Mass Index
5	Age

TABLE 2: Selected Data Attributes

Based on correlation analysis and the findings of Olisah et al. [1], Zou et al. [8], and Maniruzzaman et al. [9], the following attributes were selected for model training: Glucose, Blood Pressure, Skin Thickness, BMI, and Age (Table *2*) [7,12]. These features are strong indicators of diabetes and significantly influence prediction accuracy.

Applying machine learning

For model construction, 80% of the preprocessed dataset is used for training and the remaining 20% for testing. Several machine learning techniques are applied to predict diabetes outcomes using the Pima Indian Diabetes dataset [3,8]. The goal is to evaluate model accuracy and identify influential features.

SVM: SVM is a supervised learning algorithm used for classification by constructing an optimal hyperplane that separates classes in high-dimensional space. It is effective in handling complex, non-linear data distributions [7,9].

Decision Tree: A decision tree is a supervised learning model that uses a tree-like structure to represent decision rules derived from data features. It is easy to interpret and works well with both categorical and continuous variables [10].

Random Forest: Random Forest is an ensemble learning technique that combines multiple decision trees to improve prediction accuracy and reduce overfitting. Introduced by Breiman, it performs well on medical datasets with high variability [8,10].

Model Testing: After training, the models are tested on unseen data to evaluate their generalization capability using performance metrics such as accuracy, precision, recall, and F1-score [3,4].

Model Deployment: Deployment involves integrating the trained model into a production environment, enabling real-time prediction and decision-making through APIs or healthcare information systems.

Figure *2* presents the outcome of the analysis, visually summarizing the prediction accuracy and comparative performance of the applied algorithms [5].

How to cite this article:
Shukla A, Jain G, Singh Bisht S, et al. (April 28, 2026) Enhancing Early Diabetes Detection Through Machine Learning: Analyzing Data for Accurate Risk Prediction. Cureus J Comput Sci 3 : es44389-025-00058-8. DOI https://doi.org/10.7759/s44389-025-00058-8

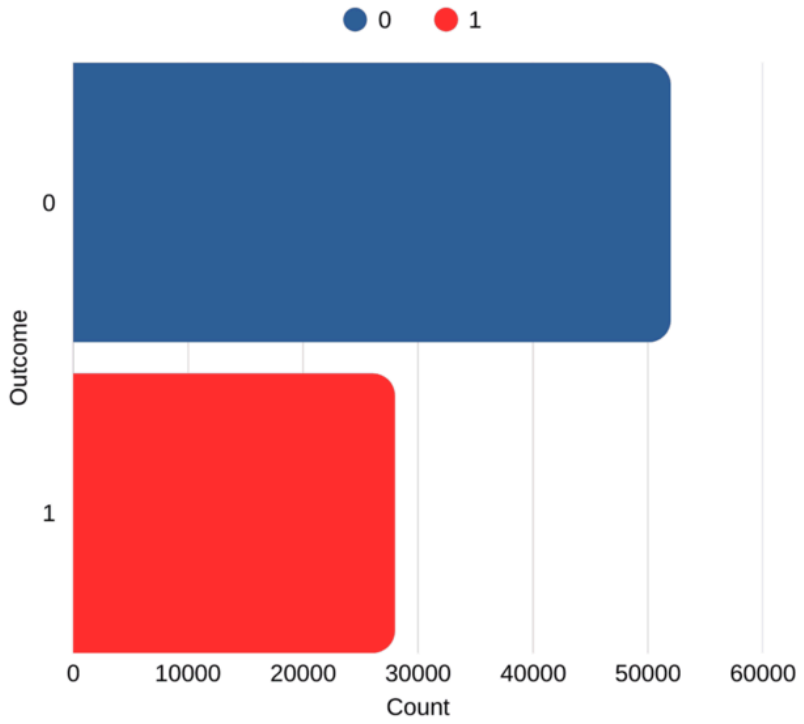

FIGURE 2: Representation of Outcome

How to cite this article:
Shukla A, Jain G, Singh Bisht S, et al. (April 28, 2026) Enhancing Early Diabetes Detection Through Machine Learning: Analyzing Data for Accurate Risk Prediction. Cureus J Comput Sci 3 : es44389-025-00058-8. DOI https://doi.org/10.7759/s44389-025-00058-8

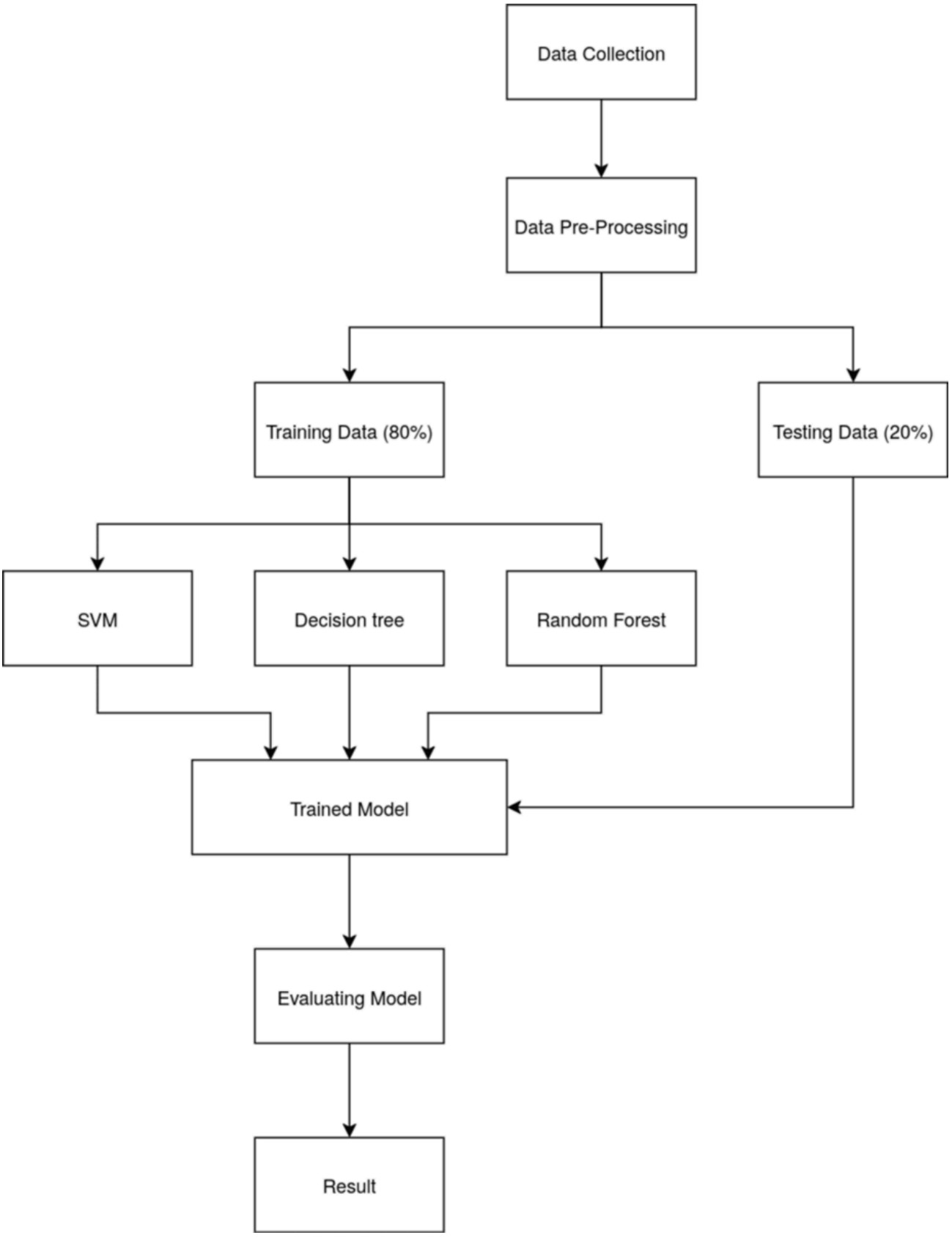

FIGURE 3: Detailed System Design

How to cite this article:
Shukla A, Jain G, Singh Bisht S, et al. (April 28, 2026) Enhancing Early Diabetes Detection Through Machine Learning: Analyzing Data for Accurate Risk Prediction. Cureus J Comput Sci 3 : es44389-025-00058-8. DOI https://doi.org/10.7759/s44389-025-00058-8

Figure *3* illustrates the detailed system design, highlighting the complete workflow from data input and preprocessing to model training, testing, and final prediction output. This architecture ensures scalability, reliability, and effective diabetes prediction.

Results And Discussion

Experimental setup

The experiments were conducted using the Pima Indian Diabetes dataset obtained from the UCI Machine Learning Repository. After preprocessing, the dataset was divided into training and testing subsets using an 80:20 split ratio. Feature selection was applied to retain clinically significant attributes, namely, Glucose, Blood Pressure, Skin Thickness, BMI, and Age. All models were trained and tested on the same data partitions to ensure a fair comparison.

Three supervised machine learning algorithms were implemented: SVM, Decision Tree, and Random Forest. The models were trained using default hyperparameters commonly adopted in related studies to maintain reproducibility and consistency. Model performance was evaluated using training accuracy and testing accuracy, as these metrics provide insight into both learning capability and generalization performance [13].

Results

The experimental results demonstrate varying levels of performance across the three machine learning algorithms. Random Forest achieved the highest predictive accuracy, with a testing accuracy of 97.10%. The SVM classifier yielded a comparatively lower testing accuracy of 79.17%, while the Decision Tree achieved a test accuracy of 90.22%. Table *3* summarizes the training and testing accuracy of all three models.

The SVM model recorded a training accuracy of 75.15% and a testing accuracy of 79.17%. This performance indicates that the SVM was able to capture general trends in the data but struggled to model complex, nonlinear relationships inherent in medical datasets.

The Decision Tree model demonstrated strong learning capability, achieving a training accuracy of 95.56%. However, its testing accuracy decreased to 90.22%, indicating that while the model learned the training data effectively, it exhibited signs of overfitting.

The Random Forest model outperformed the other classifiers, achieving a training accuracy of 97.85% and a testing accuracy of 97.10%. The close alignment between training and testing accuracy indicates robust generalization and reduced variance, making it the most reliable model among those evaluated.

Discussion

The experimental analysis reveals that ensemble-based learning methods are more effective for diabetes prediction compared to single-model classifiers. The superior performance of the Random Forest algorithm can be attributed to its ensemble nature, where multiple decision trees collectively reduce variance and mitigate overfitting. This property is particularly beneficial for healthcare datasets that often contain noise and complex feature interactions.

The Decision Tree model, although highly accurate on training data, showed a noticeable drop in testing accuracy. This behavior reflects its tendency to memorize training instances rather than learn generalized patterns, a well-known limitation of tree-based models when used independently.

The SVM algorithm demonstrated the lowest overall performance among the three models. Its sensitivity to feature scaling and kernel selection may have limited its effectiveness in capturing nonlinear relationships within the dataset. Additionally, the relatively small feature set may not fully exploit the strengths of SVM in high-dimensional spaces.

Overall, the analysis confirms that Random Forest provides the most balanced and reliable performance for diabetes prediction in this study. Its high accuracy and strong generalization capability make it a suitable choice for real-world clinical decision-support systems, where predictive reliability is critical.

How to cite this article:
Shukla A, Jain G, Singh Bisht S, et al. (April 28, 2026) Enhancing Early Diabetes Detection Through Machine Learning: Analyzing Data for Accurate Risk Prediction. Cureus J Comput Sci 3 : es44389-025-00058-8. DOI https://doi.org/10.7759/s44389-025-00058-8

Algorithms	Training Accuracy	Testing Accuracy
Support Vector Machine	75.15%	79.17%
Decision Tree	95.56%	90.22%
Random Forest	97.85%	97.10%

TABLE 3: Comparison Result of Algorithms

Conclusions

In conclusion, this research paper examined how machine learning can be used to predict diabetes, with the aim of improving early detection, prevention, and management. The study used different classification and clustering techniques to analyze a dataset that included factors such as glucose levels, blood pressure, BMI, and age. The results showed that the Random Forest algorithm performed best, with a training accuracy of 97.65% and a testing accuracy of 97.10%. This indicates that the algorithm was highly effective at predicting diabetes based on the available data. The paper also emphasized the importance of cleaning and preparing healthcare records, as missing or poor-quality data can negatively affect the performance of machine learning models. It also highlighted the value of selecting important features to make data processing more efficient. Overall, techniques such as SVM, Random Forest, and Decision Tree demonstrated strong potential in predicting diabetes risk. However, further research is needed to determine the most effective approaches for diabetes prediction and to address challenges such as data quality, privacy and security, and understanding gender differences.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Meet Thakur, Ankush Jain, Anshika Panwar, Garima Jain

Critical review of the manuscript for important intellectual content: Meet Thakur, Ankush Jain, Sovers Singh Bisht, Amita Shukla

Drafting of the manuscript: Ankush Jain, Anshika Panwar, Sovers Singh Bisht, Garima Jain

Acquisition, analysis, or interpretation of data: Sovers Singh Bisht, Amita Shukla

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

How to cite this article:
Shukla A, Jain G, Singh Bisht S, et al. (April 28, 2026) Enhancing Early Diabetes Detection Through Machine Learning: Analyzing Data for Accurate Risk Prediction. Cureus J Comput Sci 3 : es44389-025-00058-8. DOI https://doi.org/10.7759/s44389-025-00058-8

Data Availability Statements

Acknowledgements

This article was previously presented at the International Conference on NEXT GENERATION TECHNOLOGY IN SCIENCE AND ENGINEERING (ICNGTSE - 2025) at Pranveer Singh Institute of Technology, Kanpur (PSIT) on 22nd–23rd August 2025.

References

1. Olisah CC, Smith L, Smith M: Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective. Computer Methods and Programs in Biomedicine. 2022, 220:106773. 10.1016/j.cmpb.2022.106773
2. Reshmi S, Biswas SK, Boruah AN, Thounaojam DM, Purkayastha B: Diabetes prediction using machine learning analytics. 2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON), Faridabad, India. 2022, 108-112. 10.1109/COM-IT-CON54601.2022.9850922
3. Shafi S, Ansari GA: Early prediction of diabetes disease & classification of algorithms using machine learning approach. Proceedings of the International Conference on Smart Data Intelligence (ICSMDI 2021). 2021, 1-9. 10.2139/ssrn.3852590
4. Kumar PBM, Perumal RS, Nadesh RK, Arivuselva K: Type 2: diabetes mellitus prediction using deep neural networks classifier. International Journal of Cognitive Computing in Engineering. 2020, 1:55-61. 10.1016/j.ijcce.2020.10.002
5. Hassan MM, Mollick S, Yasmin F: An unsupervised cluster-based feature grouping model for early diabetes detection. Healthcare Analytics. 2022, 2:100112. 10.1016/j.health.2022.100112
6. Zhu J, Xie Q, Zheng K: An improved early detection method of type-2 diabetes mellitus using multiple classifier system. Information Sciences. 2015, 292:1-14. 10.1016/j.ins.2014.08.056
7. Ganie SM, Malik MB: An ensemble Machine Learning approach for predicting Type-II diabetes mellitus based on lifestyle indicators. Healthcare Analytics. 2022, 2:100092. 10.1016/j.health.2022.100092
8. Zou Q, Qu K, Luo Y, Yin D, Ju Y, Tang H: Predicting diabetes mellitus with machine learning techniques. Frontiers in Genetics. 2018, 9:515. 10.3389/fgene.2018.00515
9. Maniruzzaman M, Rahman MJ, Ahammed B, Abedin MM: Classification and prediction of diabetes disease using machine learning paradigm. Health Information Science and Systems. 2020, 8:7. 10.1007/s13755-019-0095-z
10. Diabetes Dataset. Kaggle. (2021). https://www.kaggle.com/datasets/jake2024/diabetes-dataset.
11. Diabetes Data Set. (2020). https://www.kaggle.com/datasets/vikasukani/diabetes-data-set.
12. Giles CR, Desurvire E: Modeling erbium-doped fiber amplifiers. Journal of Lightwave Technology. 1991, 9:271-283. 10.1109/50.65886
13. Singal M, Jain G, Sharma N, Anand S, Raza MM, Tiwari O: Enhancing cancer detection with machine learning and deep learning: a focus on breast and skin cancer. 2024 Second International Conference Computational and Characterization Techniques in Engineering & Sciences (IC3TES), Lucknow, India. 2024, 1-6. 10.1109/IC3TES62412.2024.10877274

How to cite this article:
Shukla A, Jain G, Singh Bisht S, et al. (April 28, 2026) Enhancing Early Diabetes Detection Through Machine Learning: Analyzing Data for Accurate Risk Prediction. Cureus J Comput Sci 3 : es44389-025-00058-8. DOI https://doi.org/10.7759/s44389-025-00058-8