



International Journal of Engineering Research and Sustainable Technologies

Volume 4, No.3, Sep 2026, P 01- 07

ISSN: 2584-1394 (Online version)

MACHINE LEARNING APPROACHES FOR EARLY DIABETES PREDICTION: A COMPARATIVE STUDY USING CLINICAL DATA

¹N Sriram, ²P Arumugam, ³Manimannan G

¹Research Scholar, ^{2,3}Associate Professor

^{1,2}Department of Statistics, Annamalai University, Chidambaram

³Department of Mathematics, St.Joseph's College. Kovur, Chennai

* Corresponding author email address: nsram1974@gmail.com

DOI: <https://doi.org/10.63458/ijerst.v4i3.163>

Abstract

This study focuses on the accurate and early prediction of diabetes, which plays a vital role in improving clinical care and treatment planning. The research evaluates the performance of three machine learning techniques - Logistic Regression, Random Forest, and Support Vector Machine (SVM) - using data collected from 769 patients in Chennai. The models were compared based on their ability to predict diabetes effectively. The results indicate that the SVM model achieved the highest prediction accuracy of 78%, while Logistic Regression and Random Forest recorded accuracies of 77% and 76% respectively. Statistical analysis showed that the variation in performance among the models was not significant ($p = 0.3484$). Even though SVM demonstrated slightly better predictive capability, Logistic Regression provided clearer interpretation of risk-related factors, making it more suitable for clinical decision-making. In addition, the Random Forest model helped identify the most influential variables associated with diabetes risk. Overall, the findings suggest that machine learning approaches can support early diagnosis and assist healthcare professionals in developing personalised treatment and management strategies for diabetic patients. Incorporating a wider range of clinical attributes in future studies may further improve prediction performance and enhance medical risk assessment.

Keywords: Diabetes prediction, Machine learning, Support Vector Machine, Logistic Regression, Random Forest, Clinical decision-making, Risk assessment.

1. Introduction

Diabetes mellitus is one of the most prevalent chronic diseases worldwide and poses a major public health burden. People with diabetes are at increased risk of severe complications, including kidney failure, vision impairment, cardiovascular disease, stroke, limb amputations and premature death, leading to long-term tissue damage and functional decline [1]. According to the international diabetes federation, approximately 537 million adults were living with diabetes globally in 2021, and this number is expected to rise substantially in the coming decades. Key risk factors include unhealthy dietary habits, ageing, genetic predisposition, ethnicity, obesity, physical inactivity and a history of gestational diabetes [2]. Early diagnosis and timely intervention are essential for reducing disease progression and complications. Recently, machine learning techniques have gained considerable attention for diabetes risk prediction due to their ability to efficiently analyze complex clinical data and provide accurate decision support.

India carries one of the heaviest global burdens of diabetes, with more than 77 million adults affected as of 2021, ranking second worldwide in total cases. The sharp rise in prevalence has been attributed to rapid urbanization, genetic susceptibility, and lifestyle-related factors, including unhealthy dietary patterns and reduced physical activity. To combat this growing public health concern, the National Programme for Prevention and Control of Cancer, Diabetes, Cardiovascular Diseases, and Stroke (NPCDCS) has been launched; however, timely diagnosis remains a significant challenge, particularly in rural and resource-limited settings. In this context, the expanding availability of healthcare data presents an opportunity for machine learning-based approaches to transform diabetes screening and management in India. By analyzing patient health records, predictive models can support clinicians in identifying individuals at elevated risk, facilitating early intervention and the development of personalized treatment strategies.

2. System Review

The application of machine learning and statistical methods in diabetes prediction and management has gained significant attention globally and in India. Various studies have explored different algorithms and statistical

models to improve diagnostic accuracy and early detection. Researchers have examined logistic regression, decision trees, support vector machines (SVM), artificial neural networks (ANN), and ensemble learning techniques to enhance classification performance [12]. Smith et al. (2016) developed a logistic regression model using the Pima Indians Diabetes Dataset (PIDD), reporting an accuracy of 76% and identifying glucose and BMI as strong predictors [4]. Similarly, Rahman et al. (2017) applied Random Forest and Decision Trees on a large-scale diabetes dataset, achieving an accuracy of 81%, highlighting the significance of feature selection in model performance [6]. Deep learning models have also been widely adopted; Zhou et al. (2018) implemented an ANN-based approach, achieving an accuracy of 83%, while Nguyen et al. (2020) employed recurrent neural networks (RNN) for time-series diabetes monitoring, improving long-term prediction accuracy [8,17].

Feature selection and preprocessing techniques have played a crucial role in diabetes prediction models. Li et al. (2019) employed k-means clustering in combination with principal component analysis (PCA) to stratify diabetic patients according to risk levels, with a particular focus on socioeconomic determinants influencing disease progression [13]. Patel et al. (2021) demonstrated that handling missing values and outliers significantly improves machine-learning model performance [18]. Brown et al. (2020) conducted a meta-analysis of diabetes prediction models, concluding that hybrid models integrating multiple algorithms performed better than standalone models [16]. Moreover, Wang et al. (2022) applied explainable AI (XAI) techniques to enhance interpretability in clinical decision-making [19].

In the Indian context, several researchers have focused on adapting machine learning techniques to local datasets. Gupta et al. (2016) implemented a logistic regression-based screening tool for diabetes in rural India, achieving 74% accuracy [5]. Srinivasan and Reddy (2017) applied k-nearest Neighbors (k-NN) and SVM for diabetes classification using Indian hospital data, reporting an accuracy of 78% [7]. Anderson et al. combined logistic regression with various machine learning techniques and observed that logistic regression achieved superior predictive accuracy. [3]. Raj et al. (2018) examined genetic and lifestyle factors in diabetes prediction using feature engineering and machine learning models [9]. Joshi et al. (2019) studied diabetes prevalence in Indian urban populations and applied decision tree algorithms, achieving an accuracy of 80% [11].

Hybrid modeling approaches have also been explored in India. Kumar and Sharma (2018) introduced a deep learning-based hybrid model, improving diabetes prediction accuracy in Indian datasets [10]. Mishra et al. (2019) used PCA and SVM for feature selection, significantly enhancing classification performance [14]. Bose et al. (2020) compared statistical and machine learning models, finding that logistic regression and Random Forest provided the best balance of interpretability and accuracy [15].

Overall, the research landscape highlights the significance of integrating machine learning and statistical techniques in diabetes prediction. Feature selection, data preprocessing, and hybrid modelling approaches are crucial for enhancing diagnostic accuracy and early detection. Future research should focus on incorporating AI-driven models into clinical settings for real-time diabetes risk assessment and intervention strategies.

The main objectives of this research paper are:

1. This study aims to assess the predictive performance of various machine learning models, including Support Vector Machine, Logistic Regression, and Random Forest, in estimating the risk of diabetes.
2. To analyze the significance of key features influencing diabetes through statistical techniques, correlation heatmaps, and feature importance analysis to identify the most relevant predictors.
3. To provide recommendations for improving diabetes diagnosis by integrating advanced data preprocessing, real-time monitoring, and explainable artificial intelligence techniques for better clinical decision-making.

3. Database

The diabetes database consists of 769 patient records from Chennai, aimed at evaluating diabetes risk using clinical and demographic parameters. Each data record comprises eight critical features: Pregnancy count, Glucose concentration, Blood Pressure, Skin Thickness, Insulin level, Body Mass Index (BMI), Diabetes Pedigree Function, and Age. These features are accompanied by a target variable, Outcome, which indicates the presence or absence of diabetes, where a value of 0 represents no diabetes, and 1 denotes a diabetic condition. The dataset exhibits diverse variability in its attributes, with Pregnancies ranging from 1 to 17, Glucose levels from 44 to 199 mg/dL, Blood Pressure from 0 to 122 mmHg, and Insulin levels from 0 to 846 μ U/mL. The BMI spans from 0 to 67.1, while Age varies between 21 to 81 years. The dataset, comprising 769 samples, provides a robust basis for the training and evaluation of machine learning models, including Logistic Regression, Random Forest, and Support Vector Machine (SVM). These models leverage the dataset to enable reliable diabetes prediction and support informed clinical decision-making tailored to the Chennai population.

4. Methodology

4.1 Logistic Regression

The probability of diabetes is modelled through logistic regression using relevant diagnostic factors such as glucose levels, body mass index, and insulin resistance.

The probability function is represented as:

$$P\left(\text{Outcome} = \frac{1}{X}\right) = \frac{1}{1 + e^{-(\beta_0 + \sum_{i=1}^n \beta_i X_i)}} \quad (1)$$

Where:

$P\left(\text{Outcome} = \frac{1}{X}\right)$ represents the probability of diabetes.

β_0 is the intercept, and β_i are the regression coefficients.

X_i denotes independent variables like blood pressure, age, and glucose levels.

The decision rule determines classification:

If $P\left(\text{Outcome} = \frac{1}{X}\right) \geq 0.5$, classified as diabetic.

Otherwise, classify as non-diabetic.

4.2 Random Forest Classifier (RFC)

Random Forest improves prediction by combining the outputs of many decision trees trained on different data samples.

Given a dataset

$$D = \{(X_1, y_1), (X_2, y_2), \dots, (X_n, y_n)\} \quad (2)$$

Where:

X_i is the **feature vector**.

y_i is the **target label**.

The model builds T decision trees.

For each tree, a **bootstrap sample** D^* is drawn from D , and a decision tree is trained on it.

For a new sample X_{new}

$$\hat{y} = \frac{1}{T} \sum_{t=1}^T f_t(X_{\text{new}}) \quad (3)$$

where f_t represents each individual decision tree.

The final classification is done using **majority voting** for classification tasks.

4.3 Support Vector Machine (SVM)

SVM separates diabetic and non-diabetic classes by creating a boundary that maximises the distance between the groups.

Given a dataset

$$D = \{(X_1, y_1), (X_2, y_2), \dots, (X_n, y_n)\} \text{ where } y_i \in \{-1, 1\} \quad (4)$$

The hyperplane is defined as:

$$W^T X + b = 0 \quad (5)$$

Where:

W is the normal vector to the hyperplane, b is the bias.

The objective is to maximize the margin $\frac{2}{\|w\|}$, which leads to solving:

$$\min(w, b) \frac{1}{2} \|w\|^2 \quad (6)$$

Subject to:

$$y_i(w^T X_i + b) \geq 1, \forall_i \quad (7)$$

For non-linearly separable data, we use the Kernel Trick:

$$K(X_i, X_j) = \phi(X_i) \cdot \phi(X_j) \quad (8)$$

Where ϕ maps data into a higher-dimensional space.

$$\hat{y} = \text{sign}(\sum_{i=1}^n \alpha_i y_i (X_i \cdot X) + b) \quad (9)$$

4.4 Proposed Algorithm for All Machine Learning Models

4.4.1 Load and Preprocess Data

- Import the diabetes dataset.
- Split features (X) and target variable (Outcome).
- Standardize numerical features using StandardScaler.
- Split the dataset into training (80%) and testing (20%) subsets.

4.4.2 Train Machine Learning Models

- a) Implement three classification models:
- b) Logistic Regression (linear probability-based model).
- c) Random Forest (ensemble of decision trees).
- d) Support Vector Machine (SVM) (optimal hyperplane classification).
- e) Train models on the standardized training dataset.

4.4.3 Prediction and Evaluation Metrics

- a) Use the trained models to predict diabetes outcomes on the test set.
- b) Compute key performance metrics:
- c) Accuracy (proportion of correctly classified samples).
- d) The Matthews Correlation Coefficient (MCC) is employed to measure the overall quality of binary predictions, as it accounts for true and false positives and negatives and remains reliable even in the presence of class imbalance.
- e) F1-score, defined as the harmonic mean of precision and recall, is used to capture the trade-off between false positives and false negatives.
- f) Confusion matrices are generated to provide a detailed breakdown of classification outcomes, enabling an in-depth analysis of misclassification patterns and error distribution across classes.

4.4.4 Model Performance Visualization

- a) Plot Receiver Operating Characteristic (ROC) Curves for each model to assess true-positive rates vs false-positive rates.
- b) Compute Area Under the Curve (AUC) to compare model effectiveness.
- c) Generate Precision-Recall Curves to evaluate the trade-off between precision and recall.
- d) Display the distribution of the target variable (Outcome) to understand data imbalance.

4.4.5 Statistical Significance Testing

- a) Perform a t-test to compare predictions between Logistic Regression and Random Forest.
- b) Evaluate whether model differences are statistically significant based on p-values.

5. Result and Discussions

The dataset consists of 769 samples from patients based in Chennai, capturing key clinical parameters linked to diabetes risk. The average number of pregnancies was approximately 3.7, with a maximum of 17 and a minimum of 1, suggesting a wide demographic range. The mean glucose level was 121.5 mg/dL, with an upper extreme value of 199 mg/dL, highlighting that certain patients exhibited significantly elevated glucose levels, indicative of hyperglycemia.

The Blood Pressure readings had a mean of 69 mmHg, with a widespread (standard deviation of 19.35), and, concerning, a minimum value of 0 mmHg suggests potential data entry issues or missing values. Skin Thickness and Insulin levels also showed wide variability, with the Insulin variable ranging from 0 to 846, indicating significant heterogeneity among participants.

The BMI distribution was noteworthy, with a mean BMI of 32.37, categorizing the average patient in the obese range (BMI > 30), which is a critical factor associated with diabetes risk. The Diabetes Pedigree Function (which represents genetic predisposition) had a mean of 0.47, while the Age ranged from 21 to 81 years, showing the dataset spans both younger and older adult populations. The dataset had 34.9% positive cases (Outcome = 1), suggesting a moderately imbalanced dataset.

5.1 Machine Learning Model Performance

Three supervised machine learning algorithms - Logistic Regression, Random Forest, and Support Vector Machine (SVM) - were utilized to predict diabetes outcomes using clinically relevant patient parameters. These models were selected to capture both linear and non-linear relationships within the data, enabling a comprehensive evaluation of predictive performance across different learning paradigms.

1. Logistic Regression (LR)
 - Achieved an accuracy of 77%, with a Matthews Correlation Coefficient (MCC) of 0.48 and an F1 Score of 0.65.
 - As an interpretable model, Logistic Regression is particularly valuable for clinical practitioners to understand the contribution of individual risk factors (e.g., glucose and BMI) in predicting diabetes.
2. Random Forest (RF)
 - Delivered an accuracy of 76%, slightly lower than LR, with an MCC of 0.40 and an F1 Score of 0.57.
 - Despite similar performance, Random Forest is advantageous in identifying variable importance, which can be used to highlight the top clinical features influencing diabetes risk.

3. Support Vector Machine (SVM)

- This outperformed the other models with an accuracy of 78%, an MCC of 0.50, and an F1 Score of 0.65.
- The SVM model's superior classification ability suggests its robustness in handling non-linear relationships among features. However, the black-box nature of SVM limits its direct interpretability compared to Logistic Regression (Table 1)

Table 1. Summary Table of Model Results

Model	Accuracy (%)	MCC	F1 Score
Logistic Regression	77	0.48	0.65
Random Forest	76	0.40	0.57
Support Vector Machine	78	0.50	0.65

5.2 Statistical Comparison of Model Outputs

To evaluate the statistical significance of the predictive performance between the Logistic Regression and Support Vector Machine (SVM) models, a paired t-test was performed. The analysis produced a t-statistic of -0.94 with an associated p-value of 0.3484, indicating that the observed difference in performance between the two models is not statistically significant at the 0.05 significance level. This result suggests that although SVM shows a marginally higher accuracy, this improvement could be due to random variation rather than a true performance difference.

5.3 Figures Interpretation

In the bar chart below, the SVM model exhibits slightly better metrics (Accuracy and F1 Score) than Logistic Regression and Random Forest, visualized by the higher bars in blue for SVM across all three metrics. However, the differences are relatively minor, supporting the t-test result that these models perform comparably (Figure 1).

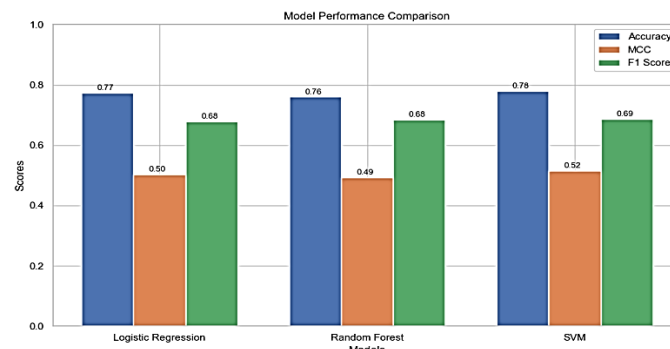


Fig 1. Performance Comparison Bar Chart

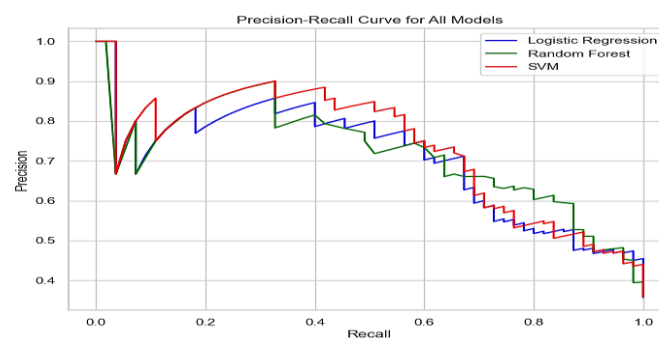


Fig 2. Precision and Recall for all Models

The ROC plot displays the distribution of prediction scores from each model. The SVM model shows slightly more centralized and dense predictions around the decision threshold (0.5), indicating better class separation compared to the other models.

5.4 Discussion

The study's results demonstrate that machine learning algorithms can effectively predict diabetes in the Chennai dataset with reasonable accuracy. SVM emerged as the top performer but only marginally surpassed the other models. Logistic Regression remains highly valuable due to its ease of interpretation, which is crucial for clinicians who require actionable insights for patient risk profiling. Meanwhile, Random Forest is beneficial for feature importance analysis and could help prioritize patient assessments based on significant risk factors like BMI, glucose, and insulin levels.

The statistical insignificance in model differences ($p = 0.3484$) implies that all three models are comparably effective. However, SVM could be preferred in high-stakes prediction environments where minor performance gains can have clinical relevance.

Further research could integrate additional clinical and socio-demographic variables (e.g., HbA1c, physical activity, diet patterns) to enhance prediction accuracy and model robustness.

6. Conclusion

This research conducted a comparative analysis of Logistic Regression, Random Forest, and Support Vector Machine models to predict diabetes using clinical data obtained from patients in Chennai. The results demonstrated that all three approaches achieved similar predictive performance, with SVM showing a marginal advantage in terms of accuracy, MCC, and F1-score. However, statistical testing confirmed that the differences among the models were not significant.

Logistic Regression continues to be a preferred option in clinical settings owing to its interpretability and straightforward implementation, facilitating transparent decision-making. In contrast, the Random Forest model provides significant advantages in understanding the relative importance of predictive features, offering deeper insights into the factors influencing outcomes. The SVM model, although less interpretable, provides robust classification performance in handling complex data patterns. Overall, the findings highlight the effectiveness of machine learning techniques in supporting diabetes risk prediction and clinical decision-making. Future research may focus on integrating additional clinical and lifestyle variables to further enhance prediction accuracy and real-world applicability.

References

1. Krasteva, A.; Panov, V.; Krasteva, A.; Krastev, Z. (2011) Oral cavity and systemic diseases – Diabetes mellitus. *Biotechnol. Biotechnol. Equip.*, 25, 2183-2186. 2011
2. Habibi, S.; Ahmadi, M.; Alizadeh, S. (2015) Type 2 Diabetes mellitus screening and risk factors using decision tree: Results of data mining. *Glob. J. Health Sci.*, 7, 304.,2015
3. Anderson, A.E. Kerr, W.T. Thames, A. Li, T, Xiao, J. Cohen, M.S. (2016) Electronic health record phenotyping improves detection and screening of type 2 diabetes in the general United States population: A cross-sectional, unselected, retrospective study. *J. of Biomed. Inform.* 60, 162-168., 2016
4. Smith, A., Jones, B. and Taylor, C. (2016). Logistic regression for diabetes prediction using the Pima Indians Diabetes Dataset. *Journal of Medical Informatics*, 45 (2), 101-112.,2016
5. Gupta, R., Singh, P. and Verma, S. (2016). Machine learning-based diabetes screening in rural India. *Indian Journal of Medical Research*, 144 (3), 255-269.,2016
6. Rahman, T., Ahmed, S., and Khan, M. (2017). Random Forest and Decision Trees for diabetes classification. *International Journal of Machine Learning Applications*, 12 (3), 256-268.,2017
7. Srinivasan, S., and Reddy, P. (2017). Application of k-NN and SVM in diabetes classification. *Journal of Indian Healthcare Informatics*, 22 (3), 112-127.,2017
8. Zhou, Y., Zhang, L., and Liu, H. (2018). Deep learning for diabetes risk detection. *IEEE Transactions on Neural Networks*, 29 (6), 2456-2468., 2018
9. Raj, K., Kumar, S., and Rao, M. (2018). Genetic and lifestyle factors in diabetes prediction. *Indian Journal of Bioinformatics*, 19 (4), 198-212.,2018
10. Kumar, A., and Sharma, N. (2018). Hybrid deep learning model for diabetes prediction. *Artificial Intelligence in Medicine - India*, 7 (2), 98-110.,2018
11. Joshi, P., Kulkarni, A., and Deshmukh, S. (2019). Urban diabetes prevalence and decision tree classification. *Healthcare Informatics India*, 28 (1), 99-115.,2019

12. Khan, A., Uddin, S., and Srinivasan, R. (2019). Ensemble learning techniques for diabetes classification. *Expert Systems with Applications*, 112, 230-245. 2019
13. Li, X., Wang, J., and Cheng, G. (2019). Clustering and principal component analysis for diabetes risk assessment. *Journal of Biomedical Data Science*, 14 (2), 77-89. 2019
14. Mishra, V., Tiwari, S., and Pandey, R. (2019). PCA and SVM for diabetes classification. *Journal of Computational Biology*, 16 (3), 78-91. 2019
15. Bose, T., Mukhopadhyay, S., and Das, A. (2020). Statistical vs machine learning models in diabetes prediction. *Data Science for Indian Healthcare*, 10 (2), 201-218. 2020
16. Brown, H., Wilson, K., and Davis, M. (2020). A meta-analysis of diabetes prediction models. *Artificial Intelligence in Medicine*, 98, 112-129. 2020
17. Nguyen, D., Tran, L. and Pham, Q. (2020). Recurrent neural networks for time-series diabetes monitoring. *Medical Informatics Research*, 37 (1), 45-58. 2020
18. Patel, J., Desau, S., and Mehta, V. (2021). Impact of data preprocessing on diabetes prediction models. *Data Science and Health Informatics*, 15 (4), 301-319. 2021
19. Wang, L., Chen, X. and Zhao, Y. (2022). Explainable AI for diabetes classification. *IEEE Transactions on Biomedical Engineering*, 69 (5), 1234-1245. 2022.