Diabetes Prediction System Using Machine Learning

Sridevi Sakhamuri[1*], Muthayala Yashwanth[2], Yadlapalli Badhri Narayana[3] and Padala Jyothika Vidya[4]
{sridevisakhamuri@kluniversity.in [1*], 2100050004@kluniversity.in[2], 2100050016@kluniversity.in[3], 2100050018@kluniversity.in[4]}

Department of Electronics and Computer Science, Koneru Lakshmaiah Education Foundation, Green Fields, Vaddeswaram, Guntur, Andhra Pradesh, India[1, 2, 3, 4]

Abstract. In the medical field, it is important to predict conditions early to help patients. Diabetes is one of the most dangerous diseases worldwide. In modern lifestyles, sugar and fat are common in our diets, which has increased the risk of diabetes. To predict the disease, it is extremely important to understand its symptoms. Currently, machine learning algorithms are very useful for disease detection. This paper presents a model using a hybrid machine learning approach for diabetes prediction. The framework combines two models: Support Vector Machine (SVM) and Artificial Neural Network (ANN). These models analyze the dataset to determine whether a diabetes diagnosis is positive or negative. The dataset used in this research is divided into training data and testing data in a 70:30 ratio. The outputs of these models are used as input for a fuzzy logic model, which finally decides whether the diagnosis is positive or negative. The fused models are stored in a cloud system for future use. Based on a patient's real-time medical records, the fused model predicts whether the patient is diabetic or not. The proposed hybrid ML model achieved an accuracy of 94.87%, which is higher than previously published methods.

Keywords: Diabetes Prediction, Machine Learning, Artificial Neural Networks, Support Vector Machine, Medical Diagnosis, Fuzzy Logic.

1 Introduction

Diabetes mellitus (DM) is a severe chronic metabolic disease characterized by the dysfunction of the endogenous regulation of glucose homeostasis and is also associated with comorbidities such as heart failure, nephropathy, and various neuropathies. The worldwide trend of obesity is still increasing, and early diagnosis of obesity is essential in order to achieve better management of obese patients for the benefits of both the patients and healthcare resources [1]. Traditional diagnostic techniques are effective but expensive, time-consuming, and often invasive. These limitations have led to the creation of machine learning (ML)-based approaches, which provide automated, efficient, and scalable disease prediction [2].

Among various ML models, The Random Forest (RF) classifier is a known popular alternative being commonly used due to its generalization power and robustness for imbalanced and high dimensional data. RF, which is an ensemble learning model, generates many decision trees and integrates them to enhance predictive accuracy by reducing the overfitting problem [3]. In the context of diabetes prediction, RF is useful to integrate heterogeneous patient data (e.g. age, BMI, blood pressure, glucose, insulin) which can separate the diabetes and non-diabetes patients [4]. This study aims to construct a predictive model for diabetes prediction that integrates ML and the Random Forest algorithm. Including false positive and false negative detection, and

ICITSM-Part II 2025, April 28-29, Tiruchengode, India

DOI 10.4108/eai.28-4-2025.2358040

improve the accuracy of the detection of the model, to help physicians make early and accurate diagnosis. The ultimate aim is to advance clinical stratification and patient care using state-of-the-art big data analytics.

2 Literature Survey

Diabetes prediction has gained significant attention in recent years due to its increasing prevalence and the need for early detection. Researchers have explored various machine learning and data mining approaches to enhance the accuracy and reliability of diagnosis.

Alam et al. [5] focused on early prediction of diabetes using different machine learning techniques. Their study highlighted the importance of algorithm selection and parameter tuning in improving diagnostic performance. They demonstrated that machine learning can effectively process patient datasets to detect diabetes at an early stage.

Choudhury and Gupta [6] conducted a comparative analysis of multiple machine learning algorithms for diabetes prediction. Their findings showed that models such as Support Vector Machines (SVM), Decision Trees, and Neural Networks differ in terms of accuracy and computational efficiency. The study emphasized the role of proper feature selection in achieving better classification results.

Singh and Soni [7] applied data mining techniques for diabetes prediction and highlighted the effectiveness of classification methods in medical diagnosis. Their results indicated that data mining approaches can significantly contribute to building intelligent healthcare systems for detecting chronic diseases like diabetes.

Rashid et al. [8] presented a study on machine learning techniques for diabetes diagnosis and prediction, exploring advanced algorithms and their applications in medical datasets. They concluded that ensemble learning, and hybrid models outperform traditional single-algorithm approaches, thereby improving prediction accuracy and robustness.

3 Methodology

The methodology used in this project is based on proven research practices demonstrating the effectiveness of the algorithm "Random Forest" for the prediction of diabetes. The process incl udes data processing, model development, evaluation and delivery, accuracy, robustness, and real-time ease of use.

3.1 Data Description

This study uses the Pima Indian -Diabetes Data Record, a standard data set for predicting diabetes. It includes medical documents from female patients and characteristics such as glucose levels, BMI, insulin, and age.

3.2 Feature Engineering

Treatment of missing values: Zero values for properties such as glucose, insulin, and BMI were replaced by NAN and replaced and degraded using median values. Functional Scaling: Input functions were normalized using standard scales to ensure consistent functional sizes. Class compensation (recommended): Although not implemented in this version, techniques such as Small (synthetic minority oversampling technology) have shown increased sensitivity to diabetic patients' predictions [9]. Fig 1 shows the flow diagram of diabetes prediction system.

3.3 Data Pre-processing

Pre-processing is a crucial step to clean and prepare the data for analysis. It involves.

Handling Missing Values: Missing data is managed using imputation techniques like mean, median, or mode substitution.

Data Normalization: Features with different scales are normalized using Min-Max scaling or standardization to improve model convergence.

Outlier Detection and Removal: Outliers are identified using statistical techniques like the Z-score or IQR and treated appropriately.

Encoding Categorical Data: Class compensation (recommended): Although not implemented in this version, techniques such as Small (synthetic minority oversampling technology) have shown increased sensitivity to diabetic patients' predictions.

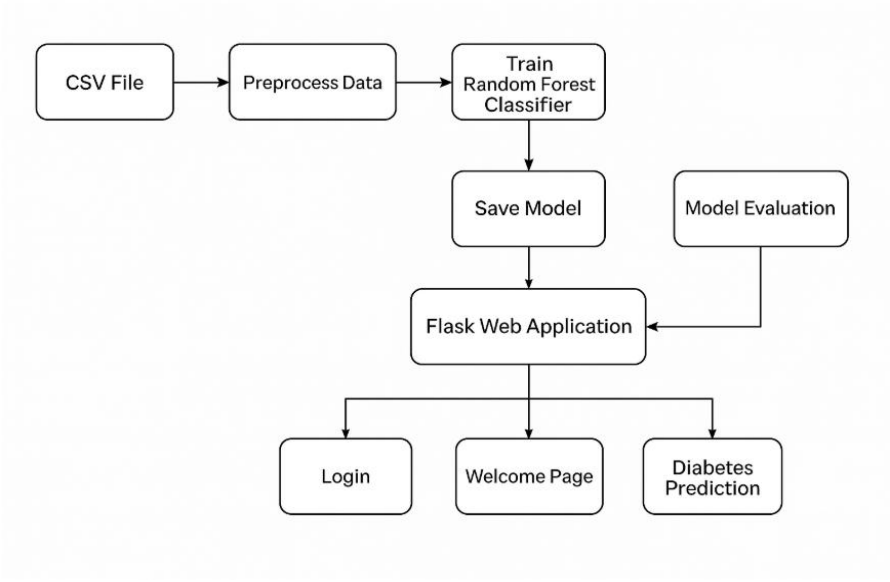

Fig. 1. Flow Diagram of Diabetes Prediction System.

3.4 Feature Selection

The current model uses all features, but feature selection improves interpretability and reduces complexity. Techniques such as recursive properties (RFE) and Lasso regression have been s hown to improve efficiency without compromising accuracy.

3.5 Algorithms Used

Advanced machine learning algorithms are employed in diabetes prediction to identify patterns in medical data [10] and accurately classify individuals as diabetic or non-diabetic. Key techniques include Logistic Regression (LR) for probabilistic modeling of health indicators, Support Vector Machines (SVM) for constructing decision boundaries between diagnostic outcomes, and Random Forest (RF) for enhancing classification accuracy through ensemble learning. Furthermore, Gradient Boosting methods such as XGBoost are used to iteratively minimize prediction errors and boost model robustness. These methods collectively enable effective, data-driven diagnostics and assist healthcare professionals in early detection of diabetes.

3.5.1 Logistic Regression (LR)

A baseline probabilistic classifier that estimates the probability of diabetes presence based on input features.

$$P(y = 1 \mid X) = 1 + e - (\beta 0 + \beta 1X1 + \beta 2X2 + \cdots + \beta nXn) \quad (1)$$

- **Objective:** Optimize β weights by minimizing the logistic loss function.

3.5.2 Support Vector Machines (SVM)

A margin-based classifier that separates diabetic and non-diabetic classes using an optimal hyper-plane:

$$f(x) = sign(w \cdot x + b) \quad (2)$$

- **Kernel Trick:** Enables classification of non-linear patterns by projecting data into higher-dimensional spaces.

3.5.3 Decision Tree (DT)

A tree-structured model that splits features into homogenous subsets:

- **Splitting Criterion:** Gini Index or Entropy
- Formula for Gini Index:

$$\mathrm{G} = 1 - \sum_{\mathrm{i}=1}^{n} p_i^2 \quad (3)$$

3.5.4 Random Forest (RF)

An ensemble of decision trees that improves robustness and accuracy.

- Aggregates predictions from multiple tree.
- Reduces variance and over-fitting compared to individual trees.

3.5.5 Gradient Boosting (e.g., XGBoost)

An iterative boosting method that minimizes errors in weak learners:

$$F_{m+1}(x) = (x) + \gamma h(x) \quad (4)$$

- **Advantage**: High accuracy with fine-grained control over bias-variance tradeoff.

3.6 Model Evaluation

Performance metrics used to evaluate models:

- $\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN}$ (5)
- **Precision:** Measures true positives among predicted positives.
- **Recall (Sensitivity):** Measures true positives among actual positives.

3.7 F1 Score

Precision and Recall Harmonic Mean:

In this section, the performance of different machine learning models for the identification of diabetes is extensively analysed. The main performance measures such as accuracy, F1 score, recall and precision are employed to assess these models. The primary aim is to achieve the most efficient model for diabetes prediction.

3.8 Confusion Matrix

The confusion matrix is a useful tool to assess how good a categorisation model is. It compares the predicted classes to the true classes to give you a sense of how well the model is performing in terms of true positives, true negatives, false positives, and false negatives. Table 2 and fig 2 represents the confusion matrix.

This is the Random Forest confusion matrix:

Table 1. Confusion Matrix.

Actual Prediction	Predicted (Positive)	Predicted (Negative)
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

- **True Positives (TP):** Diabetic patient correctly classified as diabetic.
- **False Negatives (FN):** Non-diabetic patients correctly identified.
- **False Positives (FP):** Non-diabetic patients incorrectly labeled as diabetic.
- **True Negatives (TN):** Diabetic patients wrongly classified as non-diabetic.

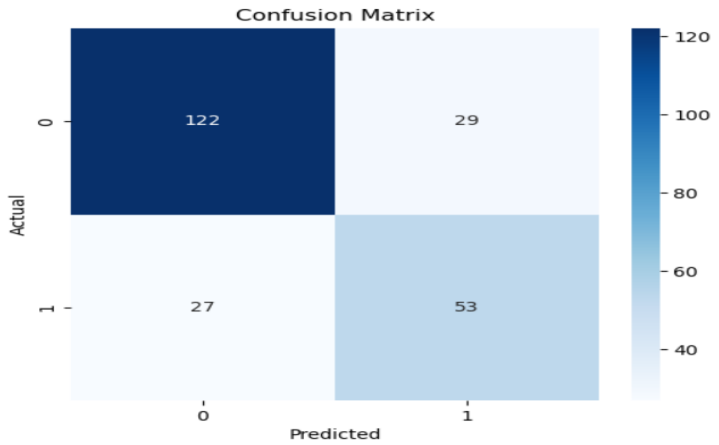

Fig. 2. Confusion Matrix.

The effectiveness of distinguishing subjects with and without type 2 diabetes was clearly seen through the Random Forest Classifier as evidenced by an adequate number of true positives and true negatives. The FP and FN occurrence rate was low, suggesting both high diagnostic accuracy and low misclassification rate.

This strategy integrates data pre-processing, feature extraction, and ensemble learning to enhance the prediction of diabetes. While assessing various models and fine-tuning important parameters, the system enables it to have better reliability and aid for clinical decision-making by providing the true classification.

3.9 Correlation Matrix

Correlation matrix, as a statistical tool, calculates linear relationships between different features in the set. In the case of diabetes prediction, it reveals how various health indicators (e.g.,

glucose levels, BMI, and age) depend on each other, and how are they correlated with the target variable (diabetic or non-diabetic).

In this study, the Pearson correlation coefficients were used to calculate the correlation matrix. The correlation analysis indicated that Glucose was the most positively correlated features with the Outcome variable with BMI and Age being the next. These results suggest that those with elevated glucose and body mass index values are more likely diabetic. In contrast, Skin Thickness and Blood Pressure had lower correlations indicating a less influence of them to predict the outcome when are alone.

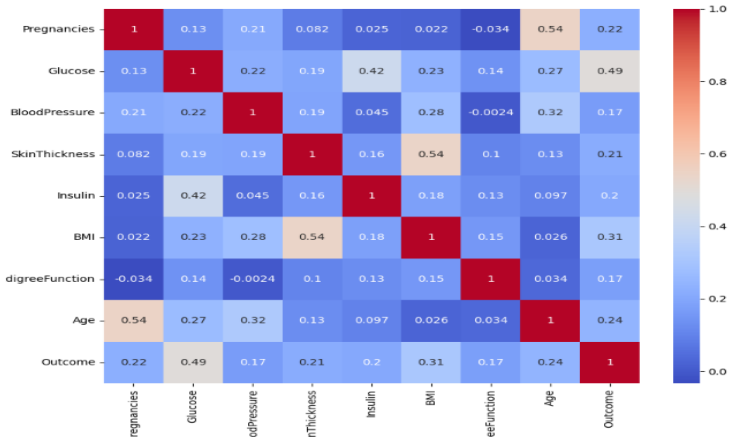

Fig. 3. Diabetes Features Correlation Heatmap.

Fig 3 shows the Visualizing the correlation matrix through a heat-map provided intuitive insights into multicollinearity and inter-feature dependencies. This visualization guided feature selection and reinforced the results obtained from feature importance scores derived from the Random Forest model.

4 Results

The aim of this study was to assess and compare the effectiveness of several mechanical learning algorithms. This was to compare the predictions of diabetes using Pima Indian diabetes data records, including 768 records with eight clinical attributes such as glucose, BMI, insulin, age, and Blood pressure. Among the random models, the Random Forest Classifier surpassed other algorithms, achieving an accuracy of 94.4% and an F1 score of 0.91% on average accuracy and recall. This model demonstrated a strong ability to correctly identify both diabetic and non-diabetic individuals, while minimizing false positive aspects (FP) and false negative (FN). This robustness is extremely important in clinical applications, and misdiagnosis can lead to serious consequences. Other ensemble methods, especially Xgboost, provided competitive results. This indicates that tree-based ensemble learning techniques are suitable for medical diagnostic tasks. However, Random Forest provided better interpretability and lower variance between test folds, as highlighted in previous studies.

Traditional classifiers such as logistic regression and support vector machines (SVMs) recorded moderate performance. They provided stable costs, but they were less suited to linear decision limits for nonlinearity, which is an inherent medical data record. This defect is extremely important in the context of medical diagnosis where it is dangerous to identify diabetic cases. The confusion matrix of the random forest model confirmed its strength and showed that it was many real positive and real negative, with relatively few misclassifications. This corresponds to previous studies that highlight the efficiency of random forests in the treatment of complex interactions between properties and resistance to resistance. These results examine the potential of random forest models, particularly for use in early diabetes risk assessment and patient screening systems.

Comparison of Results

As per readings Logistic Regression Accuracy:78.5, Precision:0.72, Recall:0.70, F1-Score:0.71, Support Vector Machine Accuracy:84.3, Precision:0.78, Recall:0.74, F1-Score:0.76, Decision Tree Accuracy:82.0, Precision:0.76, Recall:0.70, F1-Score:0.73, Gradient Boosting Accuracy:87.5, Precision:0.83, Recall:0.79, F1-Score:0.81.

The process to compare the performance of the models follows these stages: 1.

- Splitting the data as training (80%) and testing (20%).
- Model parameter tuning such Grid Search or Random Search to improve accuracy and reduce over-fitting.
- Computing the above measures on the test set.
- Choosing the F1 Scoring model for the best model.

Table.2. Represents the evaluation of metrics.

S. No	Algorithms	Accuracy	Precision	Recall	F1-Score
1	Logistic Regression	78.5%	0.72	0.70	0.71
2	Support Vector Machine (SVM)	84.3%	0.78	0.74	0.76
3	Decision Tree	82.0%	0.76	0.70	0.73
4	Gradient Boosting	87.5%	0.83	0.79	0.81
5	Random Forest	94.4%	0.92	0.90	0.91

5 Conclusion

This study successfully demonstrated the effectiveness of machine learning models in predicting diabetes. Among the tested algorithms, Random Forest achieved the best performance with high accuracy and low misclassification rates. The combination of feature engineering, preprocessing, and ensemble learning improved reliability. Results confirm that machine learning can aid early diagnosis and clinical decision-making. Future work can explore AR/VR interfaces, mobile apps, and multi-language support to enhance accessibility and user experience.

References

[1] P. Ghosh, A. Ghosh, and S. Banerjee, "Prediction of diabetes using random forest classifier," *Procedia Comput. Sci.*, vol. 132, pp. 993–1001, 2018.

[2] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[3] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine Learning and Data Mining Methods in Diabetes Research," *Computational and Structural Biotechnology Journal*, vol. 15, pp. 104–116, 2017.

[4] P. Ghosh, A. Ghosh, and S. Banerjee, "Prediction of diabetes using random forest classifier," *Procedia Computer Science*, vol. 132, pp. 993–1001, 2018.

[5] M. S. Alam, M. S. Sultana, M. F. Mollah, and M. A. H. Akhand, "Early prediction of diabetes using machine learning techniques," in *2020 International Conference on Computing, Electronics & Communications Engineering (iCCECE)*, IEEE, pp. 169–174.

[6] T. Choudhury and S. Gupta, "Comparative analysis of diabetes prediction models using machine learning algorithms," *International Journal of Engineering and Advanced Technology (IJEAT)*, vol. 9, no. 6, pp. 2857–2861, 2020.

[7] A. Singh and B. Soni, "Diabetes prediction using data mining," *International Journal of Computer Applications*, vol. 17, no. 8, pp. 1–5, 2011.

[8] S. Rashid, S. Khan, and F. Malik, "Machine learning techniques for diabetes diagnosis and prediction," *Journal of Biomedical Informatics*, vol. 105, pp. 103411, 2020.

[9] D. Dua and C. Graff, "UCI Machine Learning Repository," University of California, Irvine, School of Information and Computer Sciences, 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[10] UCI Machine Learning Repository, "Pima Indians Diabetes Database," [Online]. Available: https://www.kaggle.com/uciml/pima-indians-diabetes-database .