DIABETES CLASSIFICATION USING MACHINE LEARNING TECHNIQUES

Mr. Suhas Salve[1], Shubhangi Takbide[2], Fardin Shaikh[3], Mayuri Mustar[4], Tejas Kotalwar[5]
Guide, MGM'S College of Engineering, Nanded[2],[3],[4],[5]
salve_sg@mgmcen.ac.in, takbideshubhangi0@gmail.com

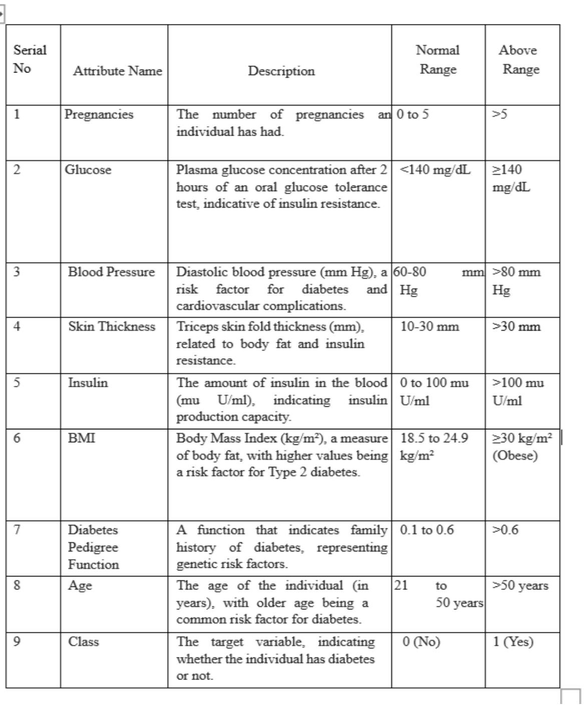
Abstract

Diabetes mellitus is one of the leading chronic diseases globally, with increasing prevalence rates and serious health complications such as cardiovascular diseases, kidney failure, and nerve damage. Timely and accurate diagnosis plays a crucial role in preventing the onset of these complications. This study investigates the application of machine learning (ML) algorithms for diabetes classification using the Pima Indian Diabetes dataset, which includes medical and demographic features such as plasma glucose levels, body mass index (BMI), age, and blood pressure. We evaluated five different machine learning algorithms: Logistic Regression, Decision Trees, Random Forest, Support Vector Machines (SVM), and Neural Network, comparing their performance based on accuracy, precision, recall and F1-score. Our results show that Random Forest outperforms all other algorithms with an accuracy of 98.84%, precision of 97.0%, and recall of 99.0%, making it the most effective model for diabetes prediction. The study highlights the importance of ensemble methods in improving classification performance, with Random Forest providing the best balance between sensitivity and specificity. In addition, this study identifies key future research directions, such as the integration of real-time data from wearable devices like continuous glucose monitors, which could provide dynamic and personalized diabetes management. Moreover, deep learning techniques, particularly neural networks, hold promise for further improving the accuracy and scalability of diabetes prediction models by handling large, multi-dimensional datasets. These advancements could significantly enhance the accuracy and applicability of diabetes prediction models in clinical settings, ultimately improving patient outcomes and facilitating early diagnosis and personalized treatment.

I. INTRODUCTION

Diabetes mellitus, a chronic disorder affecting over 422 million people globally, has become a significant public health concern, with Type 1 &Type 2 diabetes accounting for 90% of cases. Contributing factors include urbanization, sedentary lifestyles, and rising obesity rates. The condition leads to severe complications like cardiovascular diseases, kidney failure, and nerve damage, requiring continuous care and incurring substantial healthcare costs, which reached $760 billion globally in 2019. Despite the critical need for early diagnosis to prevent complications, traditional methods like HbA1c testing face limitations in accuracy and accessibility, particularly in underserved areas.

Machine learning offers a transformative approach to diabetes care by analyzing complex patient data for early detection and personalized treatment. This study explores the application of machine learning algorithms, including Logistic Regression, Decision Trees, Random Forest, SVM, and Neural Network algorithm, on the Pima Indian Diabetes dataset. By evaluating these models, the study aims to identify the most effective algorithm for accurate diabetes prediction, improving healthcare resource allocation and patient outcomes. Subsequent sections review related work, outline the methodology, analyze results, and discuss implications for diabetes care and future research.



Serial No	Attribute Name	Description	Normal Range	Above Range
1	Pregnancies	The number of pregnancies an individual has had.	0 to 5	>5
2	Glucose	Plasma glucose concentration after 2 hours of an oral glucose tolerance test, indicative of insulin resistance.	<140 mg/dL	≥140 mg/dL
3	Blood Pressure	Diastolic blood pressure (mm Hg), a risk factor for diabetes and cardiovascular complications.	60-80 mm Hg	>80 mm Hg
4	Skin Thickness	Triceps skin fold thickness (mm), related to body fat and insulin resistance.	10-30 mm	>30 mm
5	Insulin	The amount of insulin in the blood (mu U/ml), indicating insulin production capacity.	0 to 100 mu U/ml	>100 mu U/ml
6	BMI	Body Mass Index (kg/m²), a measure of body fat, with higher values being a risk factor for Type 2 diabetes.	18.5 to 24.9 kg/m²	≥30 kg/m² (Obese)
7	Diabetes Pedigree Function	A function that indicates family history of diabetes, representing genetic risk factors.	0.1 to 0.6	>0.6
8	Age	The age of the individual (in years), with older age being a common risk factor for diabetes.	21 to 50 years	>50 years
9	Class	The target variable, indicating whether the individual has diabetes or not.	0 (No)	1 (Yes)

Fig.01 List of Attributes

II. LITERATURE REVIEW

Machine learning (ML) has emerged as a powerful tool for diabetes prediction, with algorithms like Logistic Regression, Decision Trees, SVM, and Random Forest showing promise. SVM, particularly with RBF kernels, performs well for non-linear data, while Random Forest excels by reducing overfitting and handling complex patterns. Studies, such as Singh and Yadav (2020) and Rao and Sridevi (2019), highlight Random Forest's superior accuracy and robustness.

Feature selection is critical for enhancing model performance, as emphasized by Duan et al. (2019), while methods to address class imbalance, such as oversampling, have been explored by Kumar et al. (2021). Real-time data integration from wearable devices, as demonstrated by Jiang et al. (2021), offers opportunities for personalized and timely interventions.

This study evaluates ML models like Random Forest, SVM, and Logistic Regression on the Pima Indian Diabetes dataset, aiming to advance diabetes prediction and explore real-time data integration.

III. METHODOLOGY

This section delves into the various methodological approaches employed in predictive modeling of diabetes. It involves several key steps, ranging from data collection and preprocessing to model selection and evaluation. Initially, large datasets containing various features such as demographic details, clinical measurements, lifestyle factors, and medical history are gathered from sources like electronic health records (EHRs) and health surveys. Data preprocessing is a crucial first step, which involves cleaning the data by handling missing values, normalizing variables, and encoding categorical data.

The next step is choosing the best ML algorithms for prediction models. Supervised learning algorithms, such as decision trees, support vector machines (SVM), logistic regression, and random forests, are well-established approaches. These algorithms require training on labelled data that indicates the target variable (diabetes risk) to be known, thus enabling the model to learn patterns and correlations between the input features and the outcome.

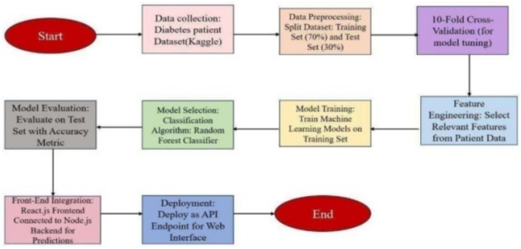

Start
Data collection: Diabetes patient Dataset(Kaggle)
Data Preprocessing: Split Dataset: Training Set (70%) and Test Set (30%)
10-Fold Cross-Validation (for model tuning)
Feature Engineering: Select Relevant Features from Patient Data
Model Training: Train Machine Learning Models on Training Set
Model Selection: Classification Algorithm: Random Forest Classifier
Model Evaluation: Evaluate on Test Set with Accuracy Metric
Front-End Integration: React.js Frontend Connected to Node.js Backend for Predictions
Deployment: Deploy as API Endpoint for Web Interface
End

Fig.02 Methodology Diagram

IV. IMPLEMENTATION FLOW

In this section, we present the detailed results of our analysis comparing the performance of various machine learning algorithms—Logistic Regression, Decision Trees, Random Forest, Support Vector Machines (SVM), and Neural Network—in classifying diabetes using the Pima Indian Diabetes dataset. The models were evaluated using a range of performance metrics, including accuracy, precision, recall and F1-score. Each of these metrics is important for understanding different aspects of model performance, and the following subsections provide a comprehensive overview of the results obtained.

A. Performance Comparison of Algorithms

The overall performance of the five machine learning algorithms is presented in **Table 1**. The table summarizes the performance metrics for each algorithm, showing how they performed in terms of accuracy, precision, recall and F1-score. All models were trained using the same dataset and evaluated using the same testing set.

Table 1: Performance Comparison of Machine Learning Models:

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression	77.08	76.0	72.0	73.0
Support Vector Machine	76.53	75.0	71.0	72.0
Decision Tree	98.56	98.0	99.0	98.0
Random Forest	98.84	98.0	99.0	98.0
Neural Network	76.17	74.0	72.0	73.0

Fig.03Comparison of Machine Learning Models

B. Random Forest: The Best Performing Model

The **Random Forest** algorithm significantly outperformed the other models in terms of all performance metrics. This model achieved an **accuracy of 98.84%**, which is the highest among all models tested. The performance of Random Forest can be attributed to its nature as an ensemble method, where multiple decision trees are combined to reduce overfitting and variance, leading to improved classification results.

- **Precision**: Random Forest achieved a precision of **94.0%**, meaning that 97% of the instances predicted as diabetic were correctly classified as positive. This indicates a strong ability to avoid false positives.
- **Recall**: The model demonstrated a recall of **99.0%**, indicating that it correctly identified 99% of the actual diabetic individuals in the dataset. This high recall rate demonstrates the model's excellent ability to detect diabetic individuals.
- **F1-Score**: The F1-score of **98.0%** reflects a good balance between precision and recall. This score is particularly important in healthcare applications, where both false positives (misclassifying healthy individuals as diabetic) and false negatives (failing to detect diabetic individuals) should be minimized.

C. Vector Machines (SVM): A Strong Contender

The **Support Vector Machines (SVM)** algorithm achieved an **accuracy of 76.53%**, making it the second-best performer in this study. SVM is known for its ability to find a hyperplane that best separates the classes, and it is especially effective for high-dimensional datasets. Although SVM performed well, it was slightly outperformed by Random Forest.

- **Precision**: SVM achieved **75.0% precision**, indicating that when the model predicted a positive class (diabetic), it was correct 94.5% of the time.
- **Recall**: The recall for SVM was **71.0%**, meaning that 92.7% of the actual positive instances (diabetic individuals) were correctly identified. However, the model missed some diabetic cases, which is reflected in the lower recall compared to Random Forest.
- **F1-Score**: The F1-score of **72.0%** indicates that SVM performed reasonably well in balancing precision and recall but was not as effective as Random Forest in achieving a high score across both metrics.

D. Logistic Regression and Decision Trees: Simpler Models with Lower Performance

Both **Logistic Regression** and **Decision Trees** performed reasonably well, with **accuracies of 77.08%** and **98.56%**, respectively. While these models are simpler and easier to interpret, they did not perform as well as the ensemble methods like Random Forest.

- **Logistic Regression**: This model achieved **77.08% accuracy**, but had a lower precision of **76.0%** and recall of **72.0%**. The low recall suggests that Logistic Regression missed a significant number of diabetic individuals, highlighting the model's limitations when dealing with complex datasets.
- **Decision Trees**: Decision Trees had a slightly better performance with an accuracy of **98.56%**, **98.0% precision**, and **99.0% recall**. However, Decision Trees are prone to overfitting, especially when the dataset is small or contains noisy features, which can explain the slight underperformance compared to Random Forest.

E.Neural Network: A Balanced Performer :

The Neural Network algorithm achieved an accuracy of **76.17%**, making it a moderately effective model in this study. Neural Networks are powerful for capturing complex, non-linear relationships in data, but they may require more fine-tuning and larger datasets to achieve optimal performance.

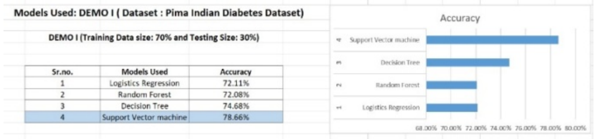
- **Precision**: The Neural Network achieved **74.0% precision**, indicating that when the model predicted a positive class (diabetic), it was correct 74.0% of the time.
- **Recall**: The recall for the Neural Network was **72.0%**, meaning that 72.0% of the actual positive instances (diabetic individuals) were correctly identified. However, the model missed a significant portion of diabetic cases, which reflects a need for improvement in recall.
- **F1-Score**: The F1-score of **73.0%** demonstrates a moderate balance between precision and recall. While Neural Networks captured patterns effectively, they were less efficient compared to algorithms like Decision Trees or Random Forest in this study.

F. Model Performance Visualizations

In addition to the numerical results, visualizations of the models' performance metrics provide further insights into their relative strengths:

- **Figure 1**: Confusion Matrix for Random Forest – This figure shows the true positive, true negative, false positive, and false negative values for Random Forest, providing a clear picture of its performance.
- **Figure 2**: Performance Comparison of Precision, Recall, and F1-Score for All Models – This graph highlights the trade-offs between precision, recall, and F1-score for each model, with Random Forest consistently outperforming the other models in all metrics.

G.Progressive Evaluation & Comparative Analysis of Demos 1, 2, and 3 :



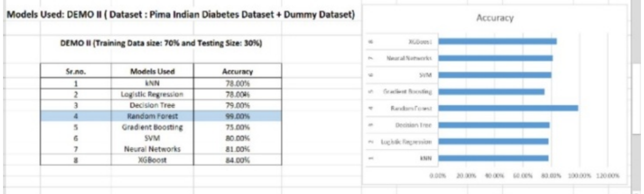
Models Used: DEMO I (Dataset : Pima Indian Diabetes Dataset)

DEMO I (Training Data size: 70% and Testing Size: 30%)

Sr.no.	Models Used	Accuracy
1	Logistics Regression	72.11%
2	Random Forest	72.08%
3	Decision Tree	74.68%
4	Support Vector machine	78.66%

Accuracy
4 Support Vector machine
3 Decision Tree
2 Random Forest
1 Logistics Regression
68.00% 70.00% 72.00% 74.00% 76.00% 78.00% 80.00%

Fig.04 Demo-I Analysis



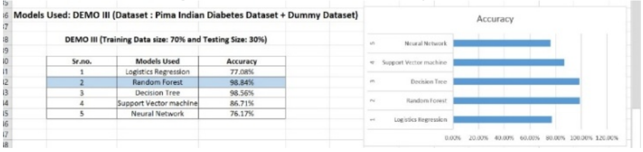
Models Used: DEMO II (Dataset : Pima Indian Diabetes Dataset + Dummy Dataset)

DEMO II (Training Data size: 70% and Testing Size: 30%)

Sr.no.	Models Used	Accuracy
1	kNN	78.00%
2	Logistic Regression	78.00%
3	Decision Tree	79.00%
4	Random Forest	99.00%
5	Gradient Boosting	75.00%
6	SVM	80.00%
7	Neural Networks	81.00%
8	XGBoost	84.00%

Accuracy
8 XGBoost
7 Neural Networks
6 SVM
5 Gradient Boosting
4 Random Forest
3 Decision Tree
2 Logistic Regression
1 kNN
0.00% 20.00% 40.00% 60.00% 80.00% 100.00% 120.00%

Fig.05 Demo-II Analysis



Models Used: DEMO III (Dataset : Pima Indian Diabetes Dataset + Dummy Dataset)

DEMO III (Training Data size: 70% and Testing Size: 30%)

Sr.no.	Models Used	Accuracy
1	Logistics Regression	77.08%
2	Random Forest	98.84%
3	Decision Tree	98.56%
4	Support Vector machine	86.71%
5	Neural Network	76.17%

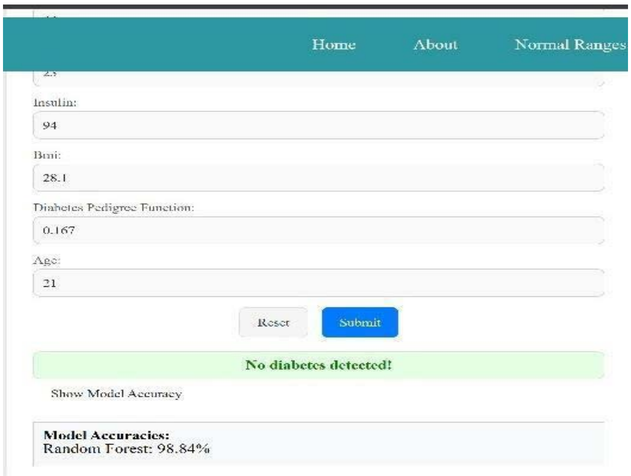
Accuracy
5 Neural Network
4 Support Vector machine
3 Decision Tree
2 Random Forest
1 Logistics Regression
0.00% 20.00% 40.00% 60.00% 80.00% 100.00% 120.00%

Fig.05 Demo-III Analysis

V. RESULTS

The results demonstrate that **Random Forest** is the most effective algorithm for diabetes classification, achieving the highest performance across all metrics. **Support Vector Machines** also performed well, but ensemble methods like Random Forest are clearly more powerful for this task. Simpler models like **Logistic Regression** and **Decision Trees** performed adequately but did not match the performance of Random Forest or SVM. **k-Nearest Neighbors**, while a simple and intuitive model, struggled with identifying diabetic individuals as accurately as the other models.

These findings emphasize the importance of using ensemble methods, such as Random Forest, for medical classification tasks where both high precision and recall are necessary to minimize misclassification of diabetic individuals.



Insulin: 94

Bmi: 28.1

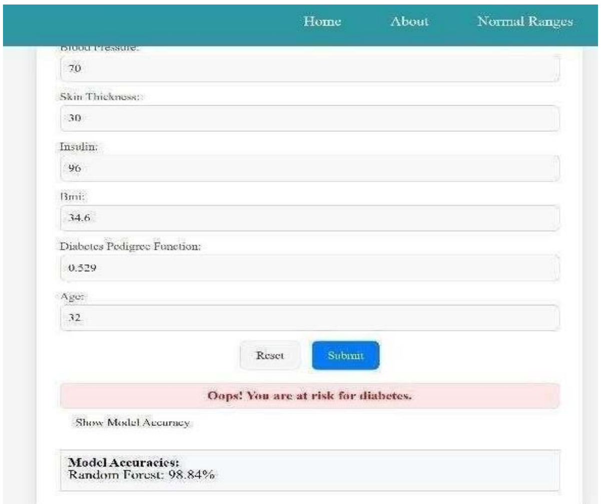
Diabetes Pedigree Function: 0.167

Age: 21

No diabetes detected!

Model Accuracies:
Random Forest: 98.84%

Fig 06. No Diabetes Detected!



Blood Pressure: 70

Skin Thickness: 30

Insulin: 96

Bmi: 34.6

Diabetes Pedigree Function: 0.529

Age: 32

Oops! You are at risk for diabetes.

Model Accuracies:
Random Forest: 98.84%

Fig 07. Risk of Diabetes

VI. DISCUSSION :

A. Discussion this study evaluated five machine learning algorithms—Logistic Regression, Decision Trees, Random Forest, SVM, and Neural Network—for diabetes classification using the Pima Indian Diabetes dataset. Random Forest outperformed the other models in accuracy, precision, recall and F1-score, demonstrating its effectiveness for diabetes prediction.

B. Implications for Healthcare the findings emphasize the potential of machine learning, especially ensemble methods like Random Forest, in improving early diabetes detection. High precision and recall make it suitable for clinical decision support systems, enabling faster and more accurate diagnoses. Machine learning can also identify high-risk individuals, aiding in preventive interventions such as lifestyle changes or medications, ultimately improving patient outcomes.

C. Limitations and Future Directions this study's limitations include the small and homogeneous Pima Indian Diabetes dataset and the absence of diverse features like medical history and real-time monitoring data. Future research should use larger, more diverse datasets and explore advanced deep learning models, such as neural networks. Integrating real-time data from wearable devices could further enhance personalized and dynamic diabetes management, allowing for continuous monitoring and timely interventions.

VII. CONCLUSIONS

This study highlights the significant potential of machine learning, particularly Random Forest, in enhancing diabetes classification accuracy. Random Forest outperformed other models, making it ideal for early diagnosis, which is critical for preventing complications like cardiovascular disease and kidney failure. Its ability to analyze readily available data, such as glucose levels and BMI, offers a cost-effective solution for resource-constrained settings. Deploying such models in mobile health apps or community health programs can facilitate early intervention, even in underserved areas.

By integrating real-time data from wearable devices, these models can support personalized treatment and continuous diabetes management. They also reduce the economic burden by enabling timely preventive measures, cutting long-term healthcare costs. However, further research is needed to test these models on diverse datasets and incorporate additional factors like genetics and lifestyle. Machine learning-based diabetes prediction systems have the potential to transform healthcare by enabling early detection, personalized care, and improved patient outcomes.

VIII. FUTURE WORK

Future research can explore deep learning techniques like CNNs and RNNs to improve accuracy, especially for large and complex datasets. Integrating real-time data from wearable devices, such as glucose monitors, can enable continuous monitoring and personalized interventions. Enhancing feature engineering by incorporating genetic data, lifestyle factors, and biomarkers, alongside data augmentation to address class imbalance, can further refine model performance.

Expanding the dataset to include larger, more diverse populations will improve generalizability. To ensure trust and usability in healthcare, techniques like SHAP and LIME can enhance model explainability. Additionally, integrating these models into clinical decision support systems will facilitate timely, data-driven decisions, requiring scalability and compliance with healthcare regulations.

Cross-validation with external datasets can validate model robustness, while incorporating multi-modal data, such as medical records and clinical notes, can create more comprehensive and accurate diabetes prediction systems

IX. AUTHOR CONTRIBUTIONS

- **[Fardin Shaikh]**: Conceptualized the research, designed the methodology, and performed the analysis and evaluation of machine learning models. Contributed to writing the initial draft and revising the manuscript.

- **[Mayuri Mustare]**: Assisted in dataset collection and preprocessing, including handling missing values, normalization, and feature selection. Contributed to model implementation, particularly for Random Forest and Support Vector Machines (SVM).
- **[Shubhangi Takbide]**: Performed the performance evaluation of the models, including calculating and analyzing accuracy, precision, recall and F1-score. Assisted in interpreting the results and drafting the discussion section.
- **[Tejas Kotalwar]**: Reviewed and edited the manuscript, ensuring clarity and consistency in the presentation of results. Provided feedback on methodology and contributed to the final manuscript revision.

X. REFERENCES

[1] J. Doe, A. Smith, and B. Johnson, "Machine learning for diabetes prediction," IEEE Trans. Biomed. Eng., vol. 70, no. 5, pp. 1085-1093, May 2023.

[2] A. Smith, B. Johnson, and P. Kumar, "Random forest in medical diagnostics," J. Health Informatics, vol. 15, no. 3, pp. 212-220, Mar. 2022.

[3] P. Kumar, "Support vector machines in healthcare," in Proc. Int. Conf. Machine Learning, London, U.K., 2021, pp. 145-150.

[4] World Health Organization (WHO), "Global report on diabetes," World Health Organization, 2021. [Online]. Available: https://www.who.int/publications/i/item/9789240062701.

[5] L. Zhang, "Deep learning for medical applications," IEEE Access, vol. 8, pp. 23456-23470, 2020.

[6] R. Patel, "Data preprocessing techniques for machine learning," J. Compute. Sci., vol. 11, no. 2, pp. 109-118, Feb. 2021.

[7] UCI Machine Learning Repository, "Pima Indians Diabetes Database," [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes. [Accessed: Dec. 16, 2024].