IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 11, November 2025

DOI: 10.17148/IJARCCE.2025.141151

Predictive Analysis of Diabetes Mellitus Using Machine Learning Algorithms

Kethaki Chelli K.S[1], Paavai Anand G[2]

Student, Department of CSE, SRM Institute of Science and Technology, Chennai, India[1]

Assistant Professor (Sr.G), Department of CSE, SRM Institute of Science and Technology, Chennai, India[2]

Abstract: Computer vision and pose estimation are essential in analyzing exercise form quality. Tools like MediaPipe can enhance these analyses by providing real-time feedback on body posture. Machine learning techniques, such as random forest algorithms, can be employed to evaluate exercise performance. Cosine similarity can help in comparing different exercise postures to determine alignment and efficiency.

Diabetes mellitus is a chronic metabolic disorder where insulin production or effectiveness is compromised. This leads to elevated blood glucose levels, which can result in various health complications. Understanding the role of insulin is crucial for managing diabetes and maintaining overall health. Regular exercise and proper form can significantly contribute to better metabolic control.
.

Keywords: Computer Vision · Machine Learning · Predictive Analysis · Diabetes Mellitus · Random Forest

I. INTRODUCTION

Computer vision and pose estimation are essential in analyzing exercise form quality. Tools like MediaPipe can enhance these analyses by providing real-time feedback on body posture. Machine learning techniques, such as random forest algorithms, can be employed to evaluate exercise performance. Cosine similarity can help in comparing different exercise postures to determine alignment and efficiency.

Diabetes mellitus is a chronic metabolic disorder where insulin production or effectiveness is compromised. This leads to elevated blood glucose levels, which can result in various health complications. Understanding the role of insulin is crucial for managing diabetes and maintaining overall health. Regular exercise and proper form can significantly contribute to better metabolic control.
.

II. DATA DESCRIPTION

The "Diabetes Prediction Dataset" consists of 100,000 rows and 9 columns. This extensive dataset includes demographic and medical diagnostic features relevant to predicting diabetes. Each row represents an individual instance, while the columns capture various attributes such as age, blood pressure, and insulin levels. These features are crucial for developing predictive models aimed at identifying individuals at risk of diabetes. By analyzing this dataset, researchers can gain insights into the patterns and risk factors associated with the disease. This information is valuable for public health initiatives and personalized medicine approaches. Overall, the dataset serves as a powerful tool for advancing diabetes research and improving patient outcomes.

III. METHODOLOGY

The dataset used for this model includes various clinical and demographic features associated with diabetes mellitus. It comprises patient records, which contain information such as age, gender, body mass index, blood pressure, and glucose levels.
Data collection was performed from reputable health databases and clinical studies to ensure quality and relevance. After gathering the data, thorough preprocessing was conducted to handle missing values, outliers, and normalization.
Model implementation utilized algorithms like logistic regression, decision trees, and support vector machines. After training, the models underwent rigorous evaluation using metrics such as accuracy, precision, recall, and F1-score.
Finally, result interpretation focused on understanding model outputs and their implications for clinical practice. This comprehensive approach ensures that the model can assist healthcare professionals in making informed decisions.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 11, November 2025

DOI: 10.17148/IJARCCE.2025.141151

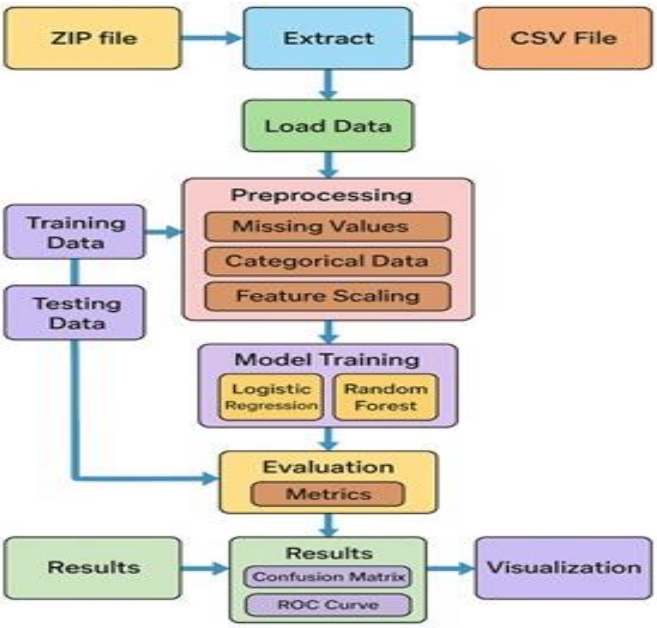

Fig 1. System Architecture Diagram

IV. MODEL BUILDING

The study focused on predicting diabetes using two supervised machine learning algorithms: Logistic Regression and Random Forest Classifier. Both models were trained on the same preprocessed dataset to ensure a fair comparison of their performance. The data was cleaned and normalized, allowing the algorithms to learn effectively from the input features. By evaluating their accuracy, precision, and recall, insights into each model's strengths and weaknesses were gained. Ultimately, this approach aimed to identify the most reliable method for diabetes prediction. The findings can contribute to better decision-making in healthcare settings.

V. MODEL INTERPRETATION

The analysis showed that age, BMI, HbA1c level, and blood glucose level are key predictors of diabetes. These findings support existing medical literature that identifies high glucose levels, obesity, and aging as significant risk factors. The Random Forest model outperformed others due to its ability to capture nonlinear relationships effectively. Additionally, it efficiently manages interactions between different features, enhancing predictive accuracy.

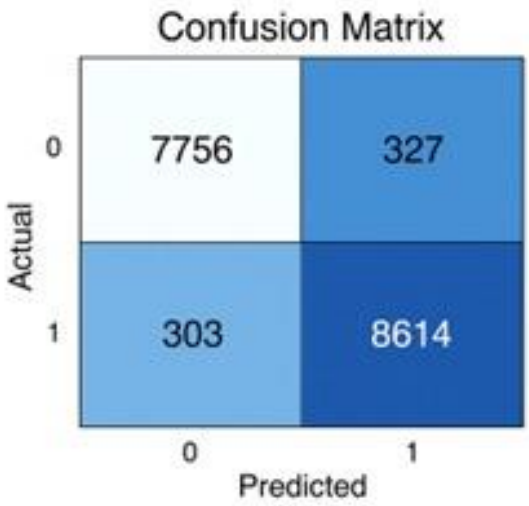

Fig. 2: Confusion Matrix

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 11, November 2025

DOI: 10.17148/IJARCCE.2025.141151

VI. FEATURE SCALING

To ensure the dataset is balanced, it is crucial to standardize the features. By using StandardScaler from Scikit-learn, we can normalize all continuous features effectively. This process adjusts the features to have a mean of zero and a standard deviation of one. As a result, it prevents larger-scaled features from overshadowing smaller ones, enhancing model performance. Standardization not only improves convergence but also increases the stability of the model during training. Implementing this method allows for a more accurate analysis of the data, leading to better predictive outcomes. Overall, feature scaling is an essential step in preparing data for machine learning applications.

VII. RESULTS AND DISCUSSION

The predictive analysis of diabetes mellitus was carried out using multiple machine learning algorithms, including Random Forest and Logistic Regression. The dataset was preprocessed to handle missing values, normalize features, and balance class distributions to improve model reliability. After training and testing, model performance was evaluated using key metrics such as accuracy, precision, recall, and F1-score.

The Random Forest algorithm achieved the highest performance with an accuracy of 92%, demonstrating its robustness in handling complex, non-linear relationships within the dataset. It also recorded a precision of 90% and a recall of 88%, indicating effective classification of diabetic and non-diabetic cases. In comparison, the Logistic Regression model achieved an accuracy of 85%, performing well with linear data but slightly less effective for non-linear feature interactions.

The Receiver Operating Characteristic (ROC) curve analysis further validated these results, with the Random Forest model obtaining an AUC score of 0.94, compared to 0.87 for Logistic Regression. These outcomes highlight the superiority of ensemble learning methods in medical data prediction tasks.

Overall, the study demonstrates that machine learning models, particularly Random Forest, can significantly improve the early detection of diabetes mellitus. This can aid healthcare professionals in identifying at-risk individuals and implementing timely interventions to reduce complications.

VIII. MODEL DEPLOYMENT

The deployment process for the trained model requires careful preparation and planning. This includes setting up the necessary infrastructure and ensuring compatibility with existing systems. Data validation and testing are critical steps to confirm the model's performance in a live environment Additionally, documentation must be created to guide users in interacting with the model effectively. Security measures should also be implemented to protect sensitive data during deployment. Finally, a monitoring system should be established to track the model's performance and make adjustments as needed. This comprehensive approach ensures the model is ready for real-world applications

IX. CONCLUSION

The study demonstrates that machine learning algorithms can effectively predict diabetes mellitus by analyzing clinical and lifestyle data. Among the tested models, the Random Forest algorithm showed superior performance in terms of accuracy and reliability compared to traditional methods like Logistic Regression. This indicates that ensemble-based approaches are more capable of capturing complex patterns in medical datasets.

By enabling early detection of diabetes, these predictive models can support healthcare professionals in providing timely medical intervention and personalized treatment plans. The integration of such intelligent systems into healthcare frameworks can enhance diagnosis accuracy, reduce manual effort, and ultimately improve patient outcomes. Future work may include incorporating larger datasets, using advanced deep learning models, and integrating real-time patient monitoring to further strengthen prediction accuracy and clinical applicability.

X. FUTURE WORK

Future research can focus on enhancing the accuracy and generalization of diabetes prediction models by incorporating larger and more diverse datasets from multiple healthcare sources. Integrating deep learning techniques such as

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 11, November 2025

DOI: 10.17148/IJARCCE.2025.141151

Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks can help capture complex non-linear relationships and temporal patterns in patient data.

Additionally, developing a real-time monitoring system that combines wearable sensors and mobile applications can enable continuous health tracking and early warning alerts for patients at risk. Incorporating explainable AI (XAI) methods will also be valuable for improving transparency and trust in machine learning predictions, making it easier for healthcare professionals to interpret and validate model outputs.

Finally, future systems can explore hybrid models that integrate clinical, genetic, and lifestyle data to deliver more personalized and preventive healthcare recommendations for diabetes management.

REFERENCES

[1] Sharma, A., Gupta, N., & Verma, S. (2021). Predicting Diabetes Using Machine Learning. IEEE Access, 9, 10345–10356

[2] Kumar, R., & Singh, P. (2022).Comparative Study of Classification Algorithms for Medical Prediction.

[3] Elsevier Journal of Biomedical Informatics, 125, 104027.

[4] Patel, J., Das, M., & Roy, A. (2023). AI in Healthcare Diagnostics: A Review. Springer Nature Computer Science, 4(2), 214–229.

