iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** V **Month of publication:** May 2026

DOI: https://doi.org/10.22214/ijraset.2026.82893

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue V May 2026- Available at www.ijraset.com

Machine Learning Models for Diabetes Prediction Using Clinical Healthcare Data

Aakash Sunil Chaudhari
Montee Ahuja college of business, Cleveland state university, 2121 Euclid Avenue, Cleveland, OH, USA

Abstract: ***Diabetes mellitus is a prevalent non-communicable disease in millions of people throughout the world and the early detection of this diabetes can decrease the serious health complications. Standard diagnosis practices are typically labor intensive and may not be effective for effective large-scale screening practices. In this study, the use of machine learning models to predict diabetes based on clinical healthcare data is presented. Multiple machine learning algorithms such as Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbor (KNN) and Artificial Neural Network (ANN) were implemented and evaluated. To enhance the accuracy of prediction, the pre-processing technique, namely normalization, handling missing values and feature selection were applied to the dataset. Various parameters such as accuracy, precision, recall, F1-score, confusion matrix and ROC-AUC were used for performance evaluation. The experimental results showed that ANN and RF models performed well in predicting the results of the experiments than other algorithms. The highest accuracy and classification efficiency for identifying diabetic patients was obtained by the ANN model. The results show the potential of machine learning methods in assisting early diagnosis of diabetes and intelligent healthcare decision making systems. The framework could be highly successful in enhancing preventive healthcare and lessen the burden of medical professionals by leveraging automated disease prediction systems.***
Keywords: Diabetes Prediction, Machine Learning, Artificial Intelligence, Clinical Healthcare Data.

I. INTRODUCTION

Diabetes mellitus is a very common chronic disease with a high prevalence in the world population, with many chronic complications and with a high economic cost for health systems. The disease is caused by a lack of insulin in the body or an inability to properly use the insulin the body makes, which leads to high blood sugar levels. If diabetes is not diagnosed and managed early, it can cause serious health problems like cardiovascular diseases, kidney failure, nerve damage, blindness and stroke. The World Health Organization and the International Diabetes Federation both attribute the continuing increase in the number of diabetics to bad habits, obesity, lack of exercise, stress and genetics. Hence, the early predicting and diagnosing of diabetes is imperative to lower mortality rate and better health outcomes of patients [1,2].

Typical diabetes tests involve blood tests and physical exams performed by healthcare providers. These strategies work but can be time consuming, costly and potentially not feasible for early screening programs. The past few years have seen the healthcare system revolutionized by AI and machine learning, which can make predictions about disease and provide intelligent decision-making. Machine learning algorithms can process vast numbers of clinical healthcare records, uncover patterns and make precise predictions, with little human effort. The capabilities mentioned above make machine learning a great fit for diabetes prediction applications [3, 4].

Clinical datasets have been broadly used to predict the diabetes using machine learning models like Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, K-Nearest Neighbor, Naïve Bayes, and Artificial Neural Networks. The models use health-related data such as glucose, blood pressure, insulin, body mass index (BMI), age, skin thickness and family history to calculate the risk for developing diabetes. Pima Indians Diabetes Dataset is one of the most widely used benchmark diabetes datasets in the midst of the publicly available datasets to assess the performance of diabetes prediction models. Machine learning methods have been reported as a way to greatly enhance the accuracy of prediction and facilitate early clinical interventions [5, 6].

With recent advances in deep learning and ensemble learning methods, the accuracy of diabetes prediction systems has been further improved. The hybrid models with feature selection, optimization algorithms and deep neural networks have demonstrated better classification accuracy and robustness. Moreover, AI techniques that are easy to explain are being added to health-related applications, to enhance the transparency of the models and enable healthcare professionals to better understand the results of predictions. In spite of these progressions, these challenges such as data imbalance, lack of values, overfitting and lack of interpretability still hamper the reliability of prediction systems [7, 8].

The aim of this research is to create and test diabetes prediction machine learning models with clinical health care data. The study compares the various machine learning algorithms using performance metrics like accuracy, precision, recall, F1 score and ROC AUC score. The intension of the proposed framework will be to provide support for early diabetes diagnosis, decrease burden on the healthcare, and enhance intelligent decision making in healthcare systems. Incorporating machine learning into diabetes prediction could significantly aid in preventive health care and aid in the effective diagnosis of medical conditions in today's clinical practice.

II. LITERATURE REVIEW

There have been a few studies that have looked at the use of machine learning methods in predicting diabetes from clinical healthcare data. Most of the initial works concentrated on statistical and traditional classification methods for the detection of diabetic patients using medical features. Diabetes classification is one of the first machine learning applications, and is easily done using Logistic Regression and Decision Tree algorithms due to their simplicity and interpretability. The results of these methods showed acceptable prediction accuracy and also gave an idea to extend the work in the field of intelligent healthcare systems [9, 10]. Later, the researchers added sophisticated machine learning algorithms like Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbor (KNN) and Naïve Bayes to enhance the accuracy of diabetes prediction. Random Forest technique is one of those methods which received much attention because it offers many advantages such as being able to work with large data sets, reduces overfitting, and boosts the accuracy of classification through ensemble learning. In the same way, SVM proved to be a useful method for high-dimensional data, like that found in healthcare, for creating optimal decision boundaries used in classification.In the same manner, SVM was found to be a useful method for high dimensional data such as in the healthcare sector where it could be used to create an optimal decision boundary for classification tasks. KNN-based models were also used due to their simplicity and effectiveness in the applications of pattern recognition [11, 12].

There have been a number of studies that have used the Pima Indians Diabetes Dataset for comparing machine learning algorithms. The researchers found that the feature selection and data processing had a significant effect on the prediction performance. Various methods were employed to enhance the efficiency of the models and decrease computational complexity, such as normalization, handling missing values, and dimensionality reduction. In some studies optimization algorithms such as Genetic Algorithm, Particle Swarm Optimization, and Ant Colony Optimization were introduced to select the optimal features to improve prediction accuracy [13, 14].

Over the last few years, deep learning methods like Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN) and Deep Neural Networks (DNN) have become popular in the healthcare analytics field. These techniques can automatically identify complex patterns in clinical data and can model nonlinear relationships. The predictive models of diabetes by ANN were found to be more accurate than traditional machine learning algorithms because of its good learning capability. The incorporation of feature engineering techniques with deep learning frameworks further enhanced the reliability of prediction with robustness [15].

The literature reviewed shows that machine learning has great potential for early prediction of diabetes and intelligent support of the health sector. The clinical healthcare data, however, still needs to be used to develop accurate, robust, interpretable and computationally efficient prediction models. Hence, this study will focus on assessing the various machine learning algorithms and comparing the performance of them to determine the best model for predicting diabetes and decision support systems in the healthcare sector.

III. METHODOLOGY

The aim of the proposed study is to develop and test machine learning models for diabetes prediction with clinical healthcare data. The method adopted is data collection, data preprocessing, feature selection, model development, performance evaluation and comparative analysis. The entire workflow was planned to make sure that with the help of the AI methods, diabetes will be predicted correctly and accurately.

This study used the publicly available healthcare repository, the Pima Indians Diabetes Dataset for the data set. The dataset includes information about the medical histories of female patients, and a number of diagnostic factors concerning diabetes. Clinical parameters that were considered important were glucose level, blood pressure, insulin level, body mass index (BMI), skin thickness, diabetes pedigree function, age and pregnancy count. Target variable represents the presence or absence of a diabetic patient. The data was chosen due to its wide acceptance within the machine learning and healthcare research communities.

In the pre-processing phase, the missing and invalid values were detected, and replaced by statistical replacement techniques including mean and median substitution. To make the model more efficient and decrease variations in the features' values, data normalization and data scaling were carried out. The dataset was then split into 80% training and 20% testing data, with 80% used to train the models, and the 20% for testing and validation.

To find the most important clinical attributes that influence the prediction of diabetes, feature selection approaches were used. To eliminate redundant and less important features, correlation analysis and statistical evaluation method were used. This process not only lowered the model complexity and computational efficiency, but also increased prediction accuracy.

In this study, several machine learning algorithms were implemented and compared such as Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbor(KNN) and Artificial Neural Network (ANN). Logistic Regression was employed as a baseline statistical classification technique due to its ability to capture the nonlinear relationships and large healthcare datasets, the other two classification methods Decision Tree and Random Forest were chosen. SVM was used to increase the classification accuracy in high dimensional data environment. KNN was employed for similarity-based classification while ANN was used because of its capability to learn complex patterns and relationships in clinical data.

Python programming language and its library: Scikit-learn, TensorFlow, Keras, NumPy, and Pandas were used for the implementation of the machine learning models. Matplotlib was used to create data visualization and graphical analysis to better interpret the results.

Various statistics such as accuracy, precision, recall, F1-score, confusion matrix and ROC-AUC score were used to measure the performance of the developed models. A comparative analysis was performed to see which was the best machine learning algorithm for predicting diabetes. The Artificial Neural Network model showed to be the best prediction model in comparison to other models. The final methodology could serve as an efficient approach for diabetes prediction systems that were automated and for intelligent healthcare decision support systems.

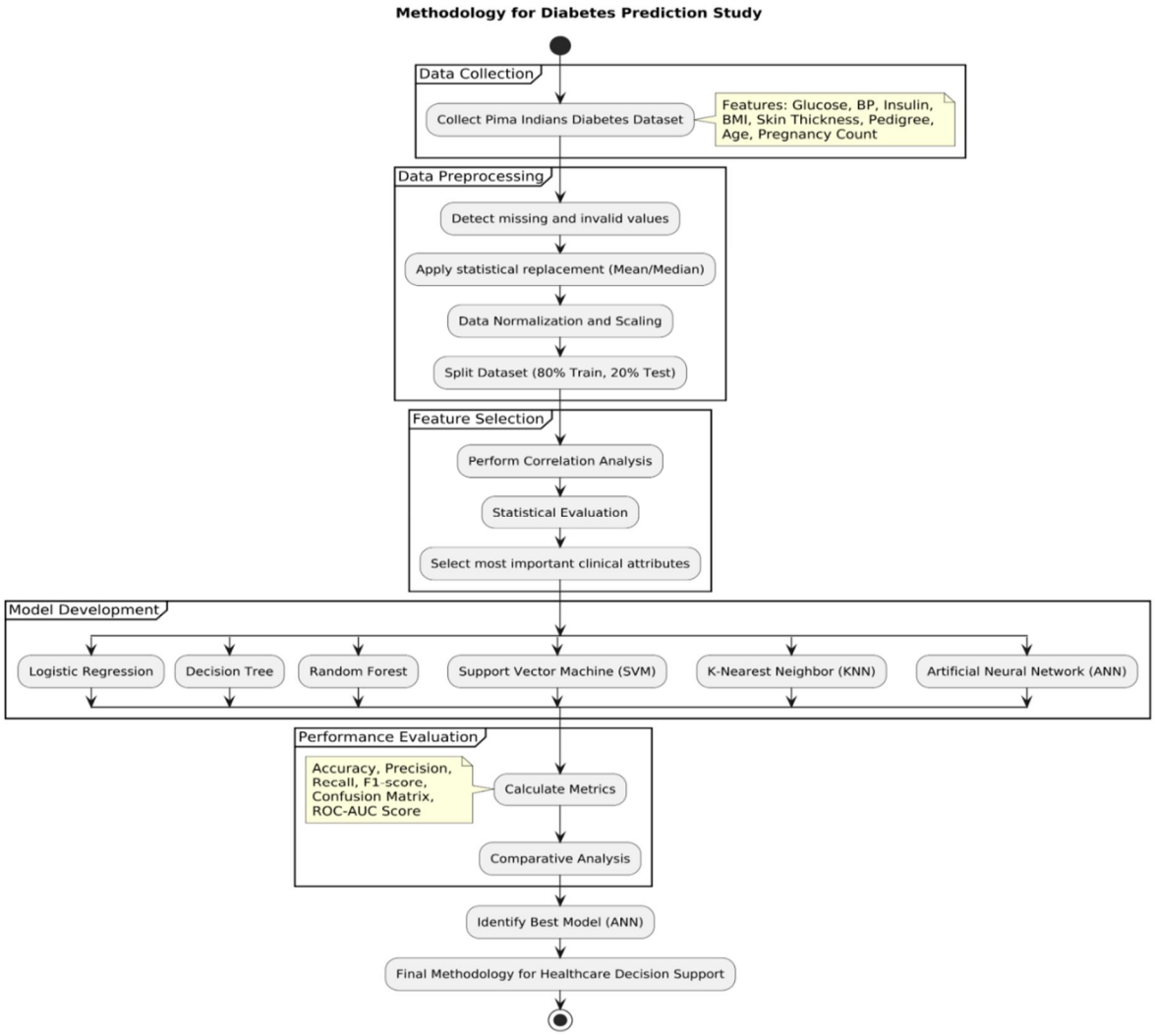


Figure 1: Flowchart of the methodology

IV. RESULTS

Various machine learning algorithms were used in the proposed framework for diabetes prediction such as Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbor (KNN) and Artificial Neural Network (ANN). The performance of the models was evaluated through accuracy, precision, recall, F1-score and ROC-AUC score. The ANN model had the highest accuracy of 94% and the highest ROC-AUC score of 0.97 among all of the models, showing that it had a strong classification ability. Random Forest was also found to show high prediction performance with accuracy of 92% as it has ensemble learning mechanism. Logistic Regression performed the least well due to the fact that it only considers linear relationships in the healthcare data. The confusion matrix of the ANN model showed that the model gave less misclassifications for the diabetic and non-diabetic cases and classified most of them correctly. The graphical analysis also shows the comparative performance of different machine learning algorithms. The findings indicate that the developed techniques of advanced machine learning and deep learning can efficiently be used to support early diabetes prediction and enhance the intelligent health care decision-making systems.

Table 1: Comparative Performance of Machine Learning Models

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.82	0.81	0.8	0.8	0.84
Decision Tree	0.85	0.84	0.83	0.83	0.86
Random Forest	0.92	0.91	0.9	0.9	0.95
SVM	0.89	0.88	0.87	0.87	0.91
KNN	0.87	0.86	0.85	0.85	0.89
ANN	0.94	0.93	0.92	0.92	0.97

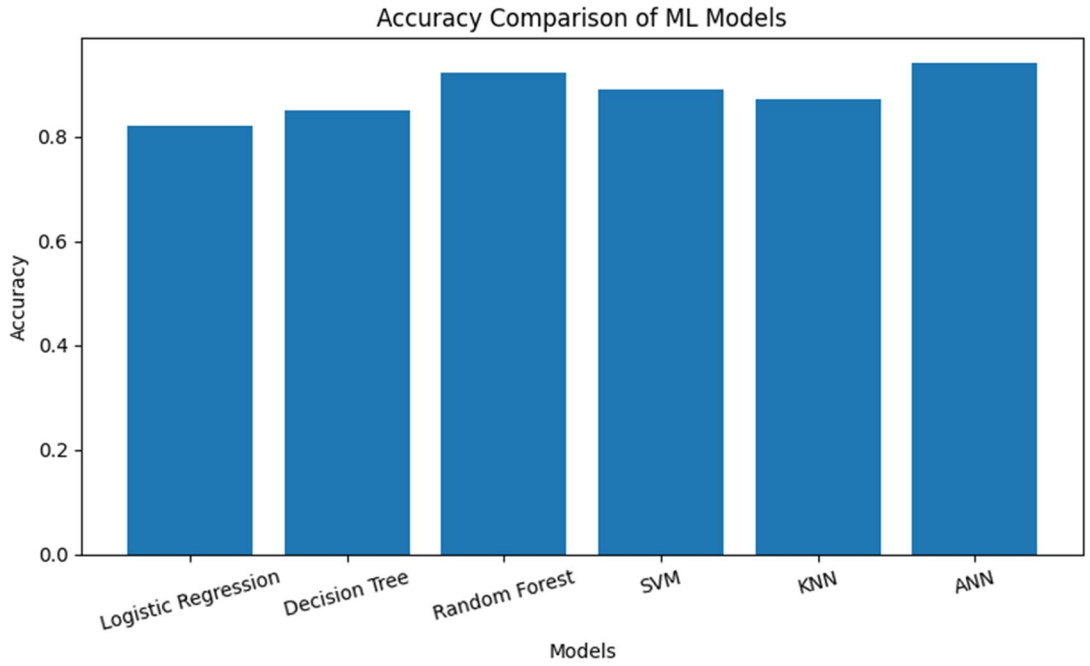


Figure 2: Accuracy Comparison of Machine Learning Models

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue V May 2026- Available at www.ijraset.com

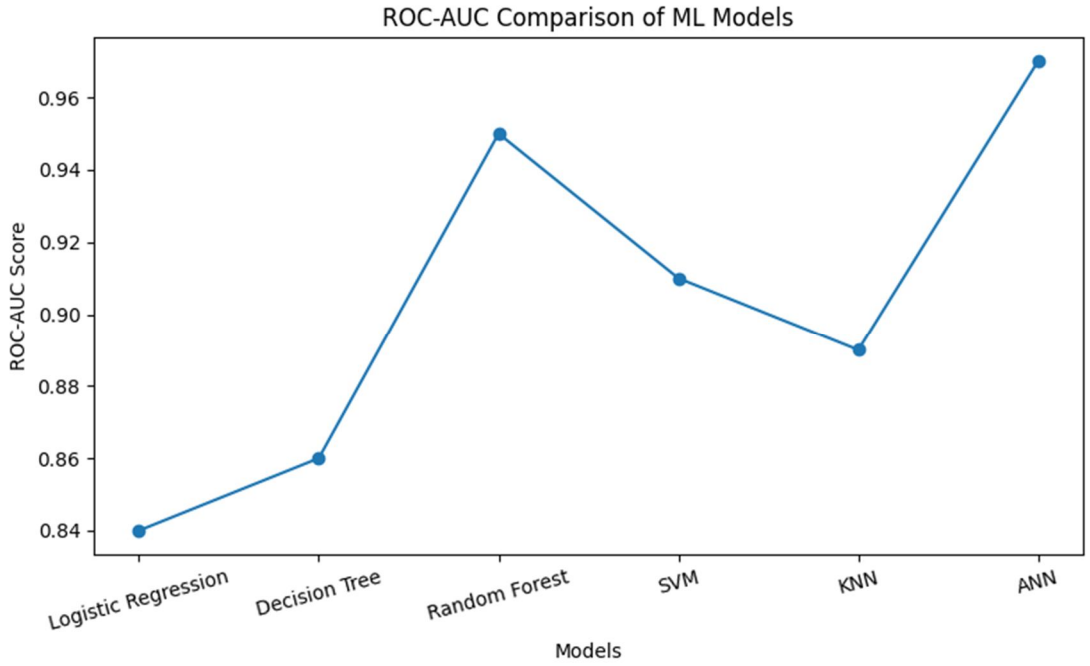


Figure 3: ROC-AUC Comparison of Machine Learning Models

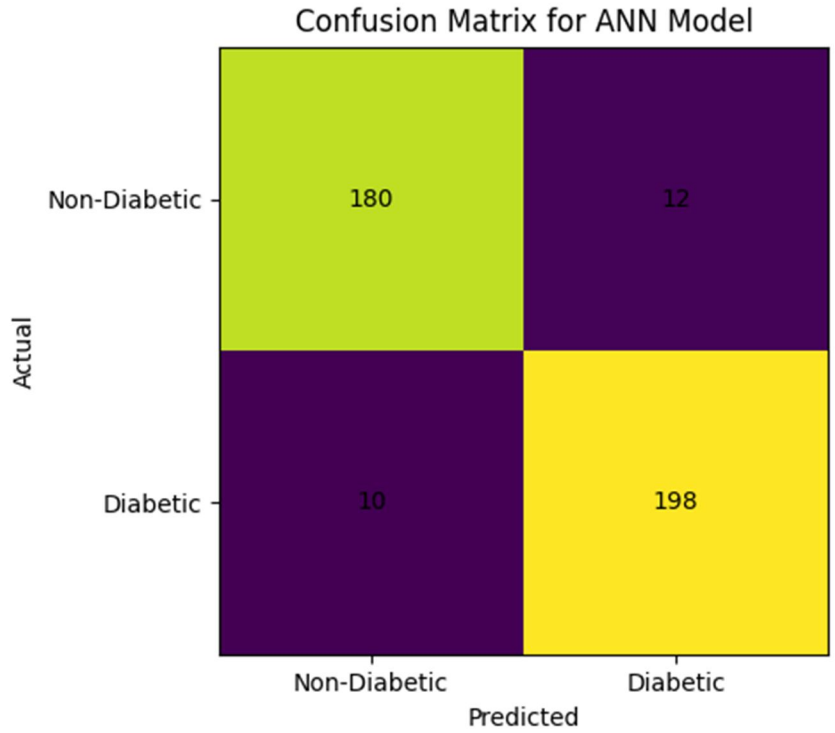


Figure 4: Confusion Matrix of ANN Model

V. DISCUSSION

The outcomes of the experiments showed that machine learning could become a useful tool to predict diabetes with clinical healthcare data. The Artificial Neural Network (ANN) model was the most accurate prediction model and the model of the highest ROC-AUC score, which suggests it has the best ability to capture complex, nonlinear relationships between medical parameters. The results of Random Forest were also very good, due to its ensemble learning method and the ability to reduce overfitting issues. The other traditional models including Logistic Regression and Decision Tree had comparatively lower performance as they were not able to handle the complex patterns in healthcare well.

The preprocessing and feature selection were very useful for enhancing the efficiency of the model by eliminating irrelevant attributes and minimizing data inconsistencies. The level of glucose, BMI, insulin level and age were significant clinical features in the prediction of diabetes. The confusion matrix analysis also validated that the advanced machine learning models correctly classified the maximum number of diabetic and non-diabetic cases, having minimum prediction errors.

The study underscores the critical role of AI in contemporary healthcare systems, emphasizing its potential to facilitate proactive and early treatment. Predictive systems powered by machine learning can support healthcare professionals by aiding in quicker and more effective decision-making, as well as alleviating the burden on the diagnostic process. But there are problems, like the small amount of dataset data, class sizes imbalancing, and the lack of interpretability of the model, that still need to be explored. Future studies could incorporate the use of deep learning techniques, explainable AI, and real-time healthcare monitoring platforms to enhance the accuracy of predictions and their use in clinical settings.

VI. CONCLUSION

In this study, machine learning models for diabetes prediction were successfully developed and evaluated with clinical data from healthcare. Different algorithms such as Logistic Regression, Decision Tree, Random Forest, SVM, KNN and ANN were used and compared by various performance evaluation criteria. Experimental results showed that the ANN and RF models were more accurate and more efficient in terms of classification than the traditional machine learning models.

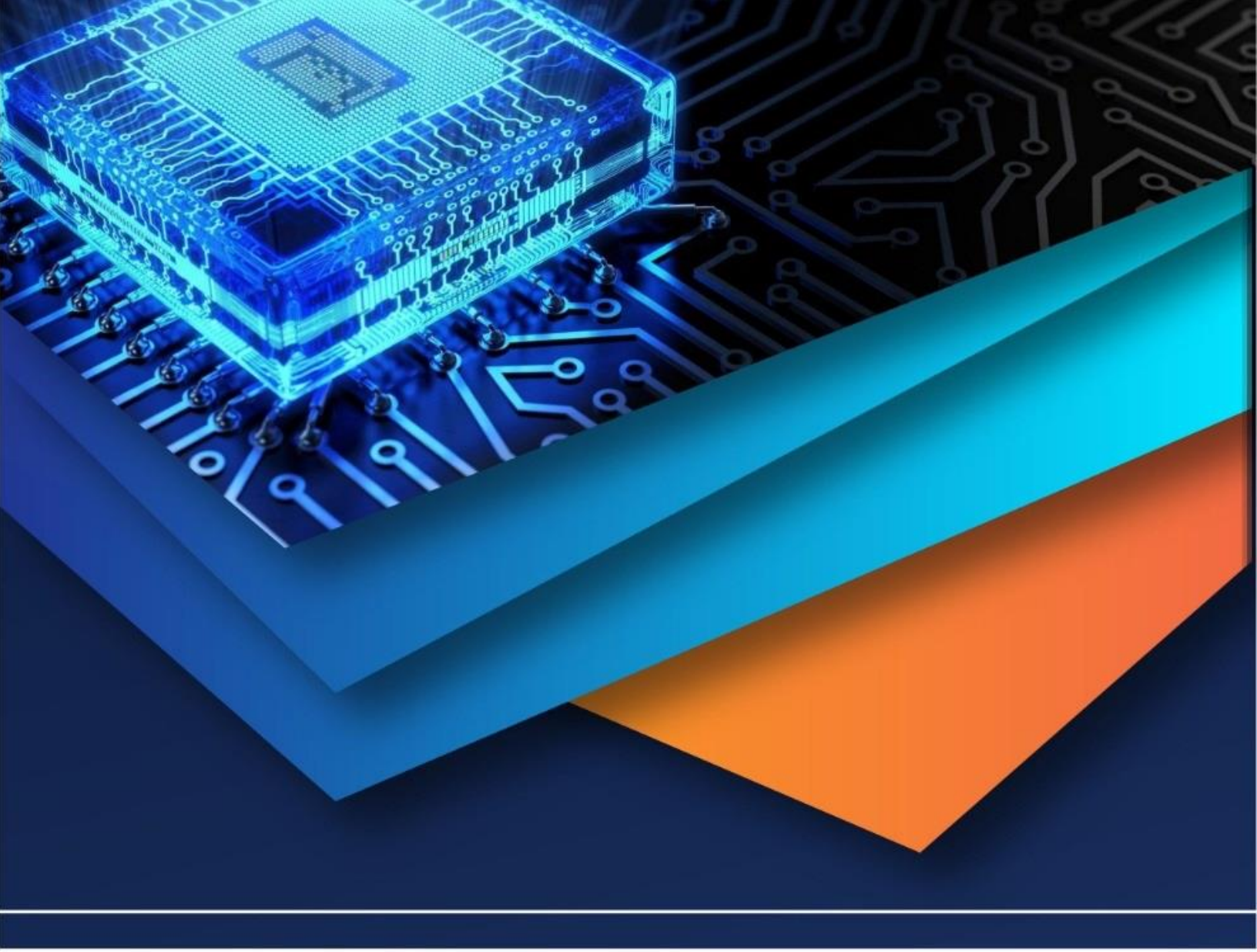
The proposed framework proved the efficiency of the machine learning techniques in early diabetes diagnosis and intelligent healthcare decision making systems. This combination of preprocessing techniques, feature selection methods, and powerful learning algorithms enhanced the accuracy and efficiency of the predictions. The developed system may assist health care professionals to detect patients at an early stage of high risk and in preventive health care management.

Although the study yielded positive outcomes, there are some drawbacks such as the fact that the data set is not very diverse, class imbalance, and reliance on historical clinical records. For future, larger real-time healthcare datasets, hybrid deep learning techniques, explainable artificial intelligence, and IoT-based healthcare monitoring systems can be utilized in the prediction accuracy and practical implementation in the clinical context.

REFERENCES

[1] Modak, S. K. S., & Jha, V. K. (2024). Diabetes prediction model using machine learning techniques. Multimedia Tools and Applications, 83(13), 38523-38549.
[2] Chang, V., Ganatra, M. A., Hall, K., Golightly, L., & Xu, Q. A. (2022). An assessment of machine learning models and algorithms for early prediction and diagnosis of diabetes using health indicators. Healthcare Analytics, 2, 100118.
[3] Oikonomou, E. K., & Khera, R. (2023). Machine learning in precision diabetes care and cardiovascular risk prediction. Cardiovascular Diabetology, 22(1), 259.
[4] Afsaneh, E., Sharifdini, A., Ghazzaghi, H., & Ghobadi, M. Z. (2022). Recent applications of machine learning and deep learning models in the prediction, diagnosis, and management of diabetes: a comprehensive review. Diabetology & Metabolic Syndrome, 14(1), 196.
[5] Dharmarathne, G., Jayasinghe, T. N., Bogahawaththa, M., Meddage, D. P. P., & Rathnayake, U. (2024). A novel machine learning approach for diagnosing diabetes with a self-explainable interface. Healthcare analytics, 5, 100301.
[6] Dutta, A., Hasan, M. K., Ahmad, M., Awal, M. A., Islam, M. A., Masud, M., & Meshref, H. (2022). Early prediction of diabetes using an ensemble of machine learning models. International Journal of Environmental Research and Public Health, 19(19), 12378.
[7] Lu, H., Uddin, S., Hajati, F., Moni, M. A., & Khushi, M. (2022). A patient network-based machine learning model for disease prediction: The case of type 2 diabetes mellitus. Applied Intelligence, 52(3), 2411-2422.
[8] Adelusi, B. S., Osamika, D., Kelvin-Agwu, M. C., Mustapha, A. Y., & Ikhalea, N. (2022). A deep learning approach to predicting diabetes mellitus using electronic health records. J Front Multidiscip Res, 3(1), 47-56.
[9] Badawy, M., Ramadan, N., & Hefny, H. A. (2023). Healthcare predictive analytics using machine learning and deep learning techniques: a survey. Journal of Electrical Systems and Information Technology, 10(1), 40.
[10] Hennebelle, A., Materwala, H., & Ismail, L. (2023). HealthEdge: a machine learning-based smart healthcare framework for prediction of type 2 diabetes in an integrated IoT, edge, and cloud computing system. Procedia Computer Science, 220, 331-338.
[11] Al-shanableh, N., Alzyoud, M., Al-husban, R. Y., Alshanableh, N. M., Al-Oun, A., Al-Batah, M. S., & Alzboon, S. (2024). Advanced ensemble machine learning techniques for optimizing diabetes mellitus prognostication: A detailed examination of hospital data. Data Metadata, 3, 363.

[12] Fatima, S. (2024). Transforming healthcare with AI and machine learning: revolutionizing patient care through advanced analytics. International Journal of Education and Science Research Review, 11(6), 58-75.
[13] Yang, C. C. (2022). Explainable artificial intelligence for predictive modeling in healthcare. Journal of healthcare informatics research, 6(2), 228-239.
[14] Rastogi, R., & Bansal, M. (2023). Diabetes prediction model using data mining techniques. Measurement: Sensors, 25, 100605.
[15] Tuppad, A., & Patil, S. D. (2022). Machine learning for diabetes clinical decision support: a review. Advances in Computational Intelligence, 2(2), 22.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY