Grenze International Journal of Engineering and Technology, June Issue

Diabetes Disease Prediction System using Machine Learning Techniques

Pratishtha Singh[1] and Arjun Singh[2]
[1-2]Greater Noida Institute of Technology /Computer Science & Engineering, Greater Noida, India
Email: rathorpratishtha12@gmail.com, innovativearjunsingh@gmail.com

***Abstract*— Nowadays diabetes mellitus is one of the most common chronic diseases worldwide, which is a impacting towards a serious situation in healthcare system. It is an important health issue and it should be managed and diagnosed at and early stage. In previous times various clinical tests were mostly used for diagnosis of diabetes which may be very time taking, costly and had a major dependency on expert interpretation. This issue had to be addressed and solved for a better treatment for those individuals who are suffering from this disease. For solving this problem diabetes disease prediction system is proposed based on machine learning techniques for the use in early detection and accurate diagnosis. This proposed system uses various factors such as body mass index, insulin level, glucose level, family history for optimized prediction. In this proposed system different types of machine learning algorithms have been implemented such as Support Vector Machine, Decision Tree, random Forest, KNN etc. for enhanced identification and effective prediction system. According to experimental results it is demonstrated that higher accuracy, robustness is achieved using ensemble-based models. This system provides a cost effective, reliable and efficient tool for better prediction and improved diagnosis of diabetes. This study highlights the importance and potential of machine learning in enhancing diabetes disease prediction.**

***Keywords*— Diabetes mellitus, Machine learning, Disease Prediction, Random Forest, Classification Algorithms.**

I. INTRODUCTION

In recent times, various diseases are spreading vastly and one of them is diabetes mellitus, it is a type of chronic disease which cannot be cured but it can be managed if diagnosed early. Millions of people are suffering from this disease and are facing major health issues. It is not good for the healthcare system in the world. Diabetes mellitus is a disorder which is caused due to higher level of glucose present in blood, which results in disturbance in insulin production or it causes lesser production of insulin [3]. This is a serious problem which must be predicted and diagnosed on time for better treatment. If it is not treated on time, it can cause various severe problems in heart disease, kidney failure and nerve issues. Hence this study majorly focuses on early detection and prediction of diabetes disease for improving disease management and treatment from serious health risks factors.

Traditionally the methods which were used for diagnosis of diabetes disease includes different lab tests and clinical expertise, these methods are very time consuming, cost effective and not had higher accuracy rate. Due to limited resources those ways for disease prediction were not much effective.

Grenze ID: 01.GIJET.12.2.65_20
© Grenze Scientific Society, 2026

Although the increasing number of medical data produced in healthcare system makes analysis in manual ways, which are more prone to human error [2,11]. These problems need to be solved and hence it is improvised with the use of machine learning techniques, for automated, intelligent systems that can be implemented by professionals for accurate prediction and diagnosis. The diagnosis of diabetes disease in early days has a crucial role in better treatment as delaying can increase complications and can also affect other body areas. The combination of machine learning based approach and algorithms makes the prediction system more enhanced with various benefits in accuracy rate, risk assessment and timely intervention [6]. Such models or system can lesser the burden on medical professionals and also this will be cost effective and scalable solution for vast amount of populations. Moreover, the prediction models can be enhanced more after easily accessible data, which leads to improved accuracy rate and reliability over time. Overall day by day the amount of increasing diabetes mellitus cases and the limitations of previously used methods for its diagnosis and prediction need to be more efficient and intelligent. By the integration of machine learning algorithms in these predictive models it can support to timely diagnosis and preventive healthcare risks. The main focus of this study is upon the development of diabetes disease prediction system using machine learning algorithms for better evaluation of disease, and enhanced model for future use. This approach will reduce burden of healthcare individuals and will reduce diagnostic complexity and had a crucial contribution in healthcare outcome.

II. Literature Review

As diabetes mellitus is increasing rapidly worldwide, the research into various enhanced and automated disease prediction system has also been in demand. Previous studies on diabetes prediction majorly relied on various rule-based approaches for analysing clinical data by the use of linear and logistic regression. These models implemented for the help in identification of important factors that causes risks such as glucose level in blood, age, and family history. However, these systems are not capable of capturing non- linear relationships, which reduces the prediction accuracy in case of complex data sets [7].
After the integration of machine learning techniques with traditional approaches, an improved system can be developed for better use.

A. Classification Models based on Machine-Learning

Machine Learning is a rapidly developing technology due to which various researchers are having a shift towards supervised learning for prediction of diabetes. For binary classifications the use of regression technique is continued because of low computational complexity [9]. The assumption of decision boundaries that are linear restricted its effectiveness. The models like decision tree gained more popularity due to their clear and concise rules and the capability of handling non-linear data very efficiently. It provided enhanced interpretability, but there were some limitations of decision tree such as they were more easily prone to issues like overfitting and are not much stable whenever trained on datasets consisting of noise.

B. Ensemble Learning Techniques

There are various ensemble learning techniques which were implemented to overcome the problems occurring because of individual classifiers such as Random Forest. It is the combination of multiple decision trees for the improvement in prediction accuracy and robustness [10]. According to various studies it is shown that the use of ensemble- based approaches enhance the model performance and provides feature importance measures, enabling optimized features understanding of risk factors.

C. Support Vector Machine Approaches

Support Vector Machine in one of the widely inspected algorithms for prediction of diabetes, the reason behind this is the capability of strong generalization and high effectiveness in higher dimensions. By the implementation of kernel functions, SVM models are able to handle complex relationships which are non-linear in nature among various features. Even so SVM also achieves higher accuracy in prediction, it has lack of interpretability surely remains a concern for clinical adoption in real world.

D. Deep Learning and hybrid approaches

In recent studies it is shown that it majorly focused on deep learning techniques such as Artificial Neural Network and other hybrid models integrating feature selection and enhancement methods with classifiers. Using these techniques has shown a effective impact as improvement in accuracy prediction by learning the large and complex feature interactions. But deep learning requirements are higher in computational resources, datasets in vast amount, and it also shows lack of transparency which makes their adoption not completely successful.

This literature review shows that every algorithm has some limitation can cannot fulfil across all dataset and conditions. While some models which are advanced also offers higher accuracy in prediction, and simpler model are helpful in providing optimized interpretability with ease of implementation. As shown in Fig. 1 the contribution of various algorithms for a optimized predictive system.

Fig 1. Machine learning algorithms used in diabetes prediction System.

III. PROPOSED METHODOLOGY

The stated methodology focuses on developing and upgrading an efficient and accurate diabetes prediction system using machine learning techniques. It involves various processes which are mentioned below:

A. Data Collection

The data set used for this study and understanding consists of patient medical records containing clinical and demographic behaviour relevant to diabetes prediction. Some common features include glucose concentration, blood pressure, body mass index (BMI), insulin level, age, skin thickness, and diabetes pedigree function. The data set may be extracted from publicly available medical reports or healthcare databases and is processed into training and testing sets for model development and estimation.

B. Data Preprocessing

Data preprocessing is a critical & crucial step to improvise model optimization. Missing or variable values in the data set are further categorised and are handled using channelized techniques such as mean or median imputation. Data interpretation or standardization is applied to ensure that all specifications are on a comparable scale. Outliers are identified and guides to reduce their impact on assumption of its accuracy [5,6]. This step deliberate data quality and reliability before model training.

C. Feature Selection and Analysis

Feature selection is the process to understand the most reliable behaviour leading to diabetes prediction. Statistical examination and correlation methodology are used to study relationships between features and the target variable [1,10]. Unessential and unneeded features are eliminated to reduce model complexity and upgrade computational effectiveness. This also helps in understanding and verifying key risk factors that can occur with diabetes.

D. Model Selection and Training

Various guided & supervised machine learning algorithms are used to create predictive prototype. These include transportation Regression, Decision Tree, Random Forest, Support Vector Machine, and K-Nearest Neighbors. Each of the models is designed and trained using the predetermined training data set. Hyperparameter tuning is carried out to enhance model performance and prevent over engineering.

E. Model Evaluation

The designed and trained models are examined using the various testing parameters to assess their desired performance. Evaluation and understanding metrics such as it's accuracy, precision, recall, F1-score, and confusion matrix are used to differentiate model effectiveness that improve reliability. Cross-examine features are applied to ensure effectiveness and generalization of the models [3,14].

F. Diabetes Prediction

Once the overall process is carried out Based on the performance results, the best-performing model is then selected for diabetes identification. The developed model predicts whether a patient is diabetic or non-diabetic based on the input data provided by medical parameters. The prediction output can be combined with a clinical understanding support system to better assist healthcare professionals in early and advance diagnosis so to carry out preventive measures.

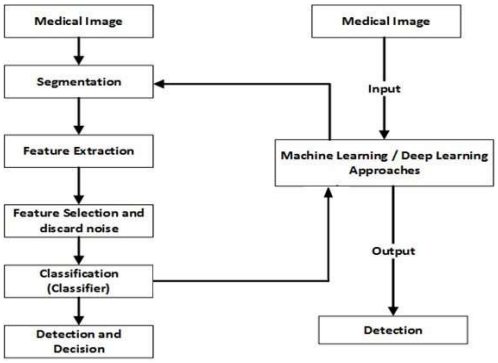

Fig. 2 Framework for diabetes disease detection using machine learning techniques.

IV. Result and Discussion

The performance of the designed diabetes identification system was examined using various machine learning algorithms after compiling data, preprocessing and specification selection. The data set was classified into training and testing sets to access the generalization accomplishments of the model [15]. Standard examination of the metrics such as accuracy, precision, recall, F1-score, and confusion matrix were used to identify and compare the effectiveness of each verifier as shown in Table 1.

TABLE I: Performance Comparison of Machine Learning Models for Diabetes Prediction

Algorithm	Accuracy	Precision	Recall	F1-Score
Decision Tree	80.1	76.9	79.8	77.6
SVM	82.3	81.7	80.2	80.9
Logistic Regression	78.5	76	75.3	76
K-NN	79.6	78.3	76.9	77.4
Random Forest	85.9	84.6	83	84.2

The results indicated that all tuned machine learning models were very capable of identifying diabetes with reasonable accuracy. Logistic lapse created stable and interpretable output although, its performance was limited when dealing with non-linear connections among clinical behaviour [4]. The decision Tree models came out with improved feasibility and transparency but were found to be delicate to fluctuations in the training data, which sometimes resulted in over fitting. Support Vector Machine resulted strongly in indicating performance, specifically in handling complex specification spaces, although its effectiveness & reliability depended heavily on careful parameters during tuning. Around all the evaluated models, Random Forest regularly generated the highest prediction precision and overall performance [11]. The marked nature of Random Forest decreased the risk of over alignment by separately acting multiple decision trees and improved effectiveness against sound in the dataset. K-Nearest Neighbors delivered satisfactory performance when applied to normalized data but that generated a higher computational cost during consumption, making it less suitable for large-scale utilisation [16].The outcome of the results highlighted that ensemble-based machine learning ways are more practical for diabetes prediction when compared to individual.

V. Conclusion

This study defines an effective diabetes prediction system by the help of machine learning methods to help early identification and preventive healthcare. By understanding clinical and demographic data, the designed system significance, the abilities of machine learning models to perfectly identify the likelihood of diabetes. Many different algorithms were introduced and evaluated to understand the most suitable way for this task.The results describe that machine learning-based models significantly performs traditional examine methods in terms of accuracy and reliability [12]. Among the diagnosis classifiers, ensemble-based methods, specifically Random Forest, generated superior performance due to their achievements to handle non-linear relationships and reduce overfitting. The involvement of such pre-defined systems into healthcare environments has the abilities to improve patient results and this help to decrease the long-term complications related with diabetes [9]. Future work may focus on incorporating larger and more diverse data sets, real-time watch, and advanced deep learning techniques to further enhance & develop prediction accuracy and clinical applicability.

FUTURE WORK

The innovation of diabetes prediction system demonstrates reliable performance; although several developments can be seen in future research to further develop its effectiveness and practical approach. One important step is the use of larger and more diverse data sets collected from multiple healthcare organisations to improve model generalization and decrease bias. In spite of considering additional clinical factors, lifestyle factors, and genetic information may also help in enhancement of prediction accuracy [8].Further studies can examine advanced deep learning methods and hybrid models that can mix with machine learning with optimization or feature selection methods to capture various complex patterns in medical data. The integration of explainable artificial intelligence (XAI) techniques can develop & improve model transparency and build trust among healthcare professionals by holding interpretable outputs. Another capability extension involves utilising the system in real-time environments by emerging it with mobile software, wearable devices, or clinical decision support systems. This would help in accessing regular watch and early risk management for individuals. Additionally, longitudinal data verification and personalized prediction models can be inspected to track disease progression and support healthcare.

REFERENCES

[1] A. Rahman, M. S. Islam, and N. Ahmed, "A machine learning approach for early prediction of diabetes using clinical data," IEEE Access, vol. 12, pp. 38421–38434, 2024.

[2] S. K. Mishra and R. Patel, "Explainable machine learning models for diabetes risk prediction," IEEE Journal of Biomedical and Health Informatics, vol. 28, no. 3, pp. 1421–1432, 2024.

[3] P. Verma, A. Singh, and R. K. Sharma, "Hybrid ensemble-based diabetes prediction system using machine learning," Journal of Healthcare Engineering, vol. 2024, pp. 1–13, 2024.

[4] M. S. Islam, M. R. Islam, and M. M. Hasan, "Diabetes prediction using machine learning techniques: A comparative study," IEEE Access, vol. 11, pp. 45621–45634, 2023.

[5] S. R. Kalyani and K. K. Swarup, "Machine learning-based early prediction of diabetes mellitus," International Journal of Medical Informatics, vol. 172, pp. 105012, 2023.

[6] A. Kumar and S. Arumugam, "An intelligent diabetes prediction system using ensemble learning," Journal of Healthcare Engineering, vol. 2022, pp. 1–12, 2022.

[7] N. Sneha and T. Gangil, "Analysis of diabetes mellitus prediction using machine learning algorithms," Procedia Computer Science, vol. 132, pp. 157–165, 2022.

[8] R. Kavakiotis et al., "Machine learning and data mining methods in diabetes research," Computational and Structural Biotechnology Journal, vol. 18, pp. 104–116, 2021.

[9] S. Perveen, M. Shahbaz, A. Guergachi, and K. Keshavjee, "Performance analysis of data mining classification techniques to predict diabetes," Procedia Computer Science, vol. 82, pp. 115–121, 2021.

[10] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2020.

[11] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2020, pp. 785–794.

[12] H. Choi, S. Kim, and J. Lee, "Prediction of diabetes using deep neural networks," IEEE Journal of Biomedical and Health Informatics, vol. 24, no. 1, pp. 210–218, 2020.

[13] M. M. Sarwar, M. I. Babar, and M. Imran, "Prediction of diabetes using artificial neural network," Proc. Int. Conf. Inf. Emerg. Technol., pp. 1–6, 2014.

[14] S. K. Rani and R. A. Kumar, "Analysis of diabetes prediction using data mining techniques," International Journal of Computer Applications, vol. 76, no. 3, pp. 1–4, 2013.

[15] P. H. Smith and R. B. Smith, "Data mining approaches for diabetes prediction," Journal of Biomedical Informatics, vol. 45, no. 2, pp. 299–307, 2012.