iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** V **Month of publication:** May 2023

DOI: https://doi.org/10.22214/ijraset.2023.52755

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 11 Issue V May 2023- Available at www.ijraset.com

Diabetes Mellitus Prediction using Machine Learning

Mangalam Chaubey[1], Rudrendra Bahadur Singh[2], Mohammad Suhaib[3], Pranjali Pal[4], KM Sejal Yadav[5]
Department of Computer Science & Engineering, Babu Banarasi Das Institute of Technology and management, Lucknow, UP, India

Abstract: ***Diabetes is a chronic metabolic disorder affecting millions of people worldwide, and machine learning has shown great potential in predicting the disease using medical and demographic features from patient data. In this paper, we propose a hybrid model of Support Vector Machines (SVM) and XGBoost for diabetes prediction, which combines the strengths of both algorithms to achieve higher accuracy and better performance. We evaluate the proposed model using the Pima Indian diabetes dataset and compare its performance with other machine learning models.***
To improve the performance of the hybrid model, we also apply feature selection techniques to select the most relevant features from the dataset. Our results show that the proposed hybrid model of SVM and XGBoost achieves higher accuracy, recall, and F1 score compared to other machine learning models such as Logistic Regression, Random Forest, and Naive Bayes. Furthermore, the performance of the hybrid model is significantly improved by feature selection, which helps to reduce the dimensionality of the dataset and focus only on the most relevant features.
Keywords: Diabetes prediction, Machine learning, Support vector machines, XGBoost, Hybrid model, Feature selection

I. INTRODUCTION

Diabetes, according to medical professionals, is a condition in which the pancreas of the human body either cannot create enough insulin (Type 1 diabetes) or the insulin that is produced cannot be utilised by the body's cells (Type 2 diabetes) [1]. Glucose is released when we eat because of the digesting process. A blood hormone called insulin encourages cells to ingest blood glucose and use it as fuel by travelling from the blood to the cells. Because cells cannot take up glucose when the pancreas is producing insufficient insulin, the glucose stays in the circulation. As a result, the blood's blood glucose/blood sugar levels rise to an extremely undesirable level. The human body experiences several symptoms of high blood sugar, including acute hunger, intense thirst, and frequent urine. The typical range of glucose concentrations in an adult is 70 to 99 mg/dL. Diabetes is indicated if the glucose level is greater than 126 mg/dl. If a person's body glucose level is between 100 and 125 mg/dl, they are said to have prediabetes. If Peer review under responsibility of The Korean Institute of Communications and Information Sciences (KICS). When blood sugar levels in the body reach dangerously high levels, heart disease, renal failure, stroke, and nerve damage may result Diabetes cannot be cured permanently. Emerging technologies in the modern world include big data, the internet of things, artificial intelligence (AI), machine learning (ML), and deep learning (DL) [2]. Patients may readily verify their health early on with the aid of ML, and it will also assist practitioners with future research [6]. It can be utilised for both regression and classification issues. Diabetes Prediction is a classification problem, thus we can group people according to whether they have diabetes or not [5]. For analysing and synthesising the data into relevant information from diverse angles, many ML approaches are helpful. Pre-processing the dataset, feature selection and extraction, training and testing, and additional evaluation are some of the procedures involved in ML.

For effective diabetes prediction a variety of ML algorithms can be used, including Support Vector Machine (SVM), Decision Tree (DT), Neural Network (NN), and Random Forest (RF). These models are trained, then tested using test datasets to determine whether the model is functioning properly.

The objectives of this research paper are threefold: firstly, to evaluate the performance of the hybrid model of SVM and XGBoost for diabetes mellitus prediction; secondly, to compare the hybrid model with other state-of-the-art machine learning models using the Pima Indian Diabetes dataset; and thirdly, to provide insights into the applicability and potential of advanced machine learning techniques in clinical settings. By undertaking this research, we aim to contribute to the growing body of knowledge on diabetes mellitus prediction using machine learning. The outcomes of this study can facilitate informed decision-making in healthcare, enable early detection of diabetes, and support preventive interventions. Furthermore, the findings may pave the way for the development of personalized healthcare strategies and the integration of machine learning into routine clinical practice.

II. LITERATURE REVIEW

Recent B Rajesh and Sangeetha [3] proposed a system in which data mining was used for classification of diabetes to determine if the person has diabetes or not. The dataset used was PIDD . In this study ten classification algorithm was involved and results were compared among them .

Anuja and Chitra [4] had proposed a system using support vector machine. SVM is a universal algorithm based on definite risk bounds of statistical learning theory so it is called structural risk Anuja and Chitra minimization principle

Raja Krishnamoorthi and Shubham Joshi [5] developed a framework to develop and assess the decision tree- based random forest and support vector machine algorithm for prediction of diabetes. The proposed work by them gave an 83% accuracy with minimum error rate.

Kamrul Hasan and Ashraful Alam [6] used a different methodology ,they developed an ensemble of machine learning classifiers. They proposed a robust framework for diabetes prediction using different ML classifiers . In this the key role was played by improving the quality of dataset where outlier rejection and filling missing values was taken care of by the authors. Such processing of data improved the kurtosis and skewness of the attribute distribution in PID dataset

Several studies have explored the use of machine learning algorithms for diabetes prediction using the Pima Indians dataset. Sharafoddini et al. (2018) used logistic regression, decision tree, and support vector machine algorithms to predict diabetes in the Pima Indians population, achieving an accuracy of up to 80%. Gupta et al. (2018) compared six different machine learning algorithms and found that the random forest algorithm performed the best, with an accuracy of 78.5%.e

III.METHODOLOGY

The problem identified is Diabetes prediction using machine learning. While reading the research papers we have identified multiple algorithms and we have compiled the accuracy score and recall score of different machine learning algorithm . While coming to solution of the problem we tried different machine learning algorithm like SVM, Logistic Regression ,etc. then we moved to ensemble techniques and finally we came to hybrid model for the prediction in our project. The methodology I described as follows

A. Data Set

Pima Indian Diabetes Database is the used data set for the prediction of diabetes It has 768 patients data with 9 attributes which includes columns like glucose, pregnancies, skin thickness etc. The outcome predicts whether the person has diabetes or not.

The Pima Indian Diabetes dataset is a widely used dataset in diabetes prediction research. It originated from a study conducted on female individuals of Pima Indian heritage in Arizona, United States. The dataset contains several medical and demographic attributes, including age, number of pregnancies, body mass index (BMI), blood pressure, skin thickness, insulin level, and diabetes pedigree function. These attributes serve as potential predictors for the presence or absence of diabetes.

In this study, the Pima Indian Diabetes dataset was obtained from a reliable source and carefully examined to understand its structure and variables. Detailed descriptions of the attributes, their data types, and any missing values were documented. This dataset serves as the foundation for subsequent preprocessing steps and model development

B. Preprocessing

In preprocessing we have used the Standard Scaler function to standardized the input data and then the training test split was 80:20. Prior to model development, thorough preprocessing steps were implemented to ensure data quality and improve the performance of the prediction models. Data cleaning involved handling missing values, outliers, and erroneous entries. Missing values were either imputed using appropriate techniques such as mean imputation or were excluded from the analysis, depending on the extent of missingness and the nature of the variable.

Feature engineering techniques were applied to derive additional meaningful features from the existing attributes. This involved creating new variables based on domain knowledge and medical insights. For instance, the body mass index (BMI) and blood pressure attributes could be combined to create an indicator of metabolic health. Feature scaling was also performed to normalize the variables, ensuring that each attribute had a similar range and distribution.

C. Model Building

The hybrid model of Support Vector Machines (SVM) and XGBoost was developed and implemented for diabetes prediction. SVM is a powerful classification algorithm that seeks to find an optimal hyperplane to separate the classes. XGBoost, on the other hand, is a gradient boosting framework known for its ability to handle complex data patterns and achieve high predictive accuracy.

To implement the hybrid model, SVM and XGBoost algorithms were combined using an ensemble approach. This involved training both models individually on the pre-processed dataset and then aggregating their predictions using a suitable strategy, such as voting or averaging. The hybrid model utilized the strengths of both algorithms, aiming to improve prediction performance compared to using either model individually.

D. Performance Evaluation

To evaluate the performance of the hybrid model, a comparative analysis was conducted with other machine learning models. Several popular algorithms, including logistic regression, decision trees, random forests, and neural networks, were implemented and trained on the same preprocessed dataset. The models were assessed based on performance metrics such as accuracy, recall, precision, and F1-score.

The comparative analysis involved training each model using the same training dataset, tuning their respective hyperparameters using suitable techniques like cross-validation, and evaluating their performance on a separate test dataset. The results obtained from each model were compared, and statistical tests, such as t-tests or ANOVA, may have been conducted to determine the significance of any observed differences in performance.

IV. RESULT

The performance of the hybrid model, consisting of SVM and XGBoost, was compared to other machine learning models, including logistic regression, decision trees, random forests, and neural networks. The evaluation focused on metrics such as accuracy, recall, precision, and F1-score. The results were analysed to determine whether the hybrid model outperformed or achieved comparable performance to the other models.

The accuracy metric indicates the overall correctness of the model's predictions, representing the proportion of correctly classified instances. Recall, also known as sensitivity, measures the model's ability to correctly identify positive instances, such as predicting the presence of diabetes in this case. Precision reflects the proportion of correctly predicted positive instances among all instances predicted as positive. F1-score is a harmonic mean of precision and recall, providing a balanced measure of the model's performance.

1) The results show that the SVM+XGBoost hybrid model achieves the highest accuracy of 0.800 and recall of 0.746, outperforming all other models.

2) The decision tree+XGBoost hybrid model achieves an accuracy of 0.759 and recall of 0.694, also outperforming all other models except for SVM+XGBoost.

3) The SVM model achieves an accuracy of 0.785 and recall of 0.672, which is lower than the SVM+XGBoost hybrid model.

Table -1: Model Comparison

Model	Accuracy	Recall
Logistic regression	72.2-80.4%	49-84.4%
Decision Tree	74.3-79.4%	60.4-77.1%
Support Vector Machine	75-86.5%	65.8-80.7%
Random Forest	67-78.5%	47.5-80.0%
Point K nearest Neighbors	64-78.9%	47.5-80.0%
Artificial Neural network	73.3-84.2%	63-84.4%
Ensemble Model	77.6-82.5%	71.4-78.6%

4) The XGBoost model achieves an accuracy of 0.778 and recall of 0.670, which is lower than both the SVM+XGBoost and decision tree+XGBoost hybrid models. The decision tree, random forest, logistic regression, and naive Bayes models achieve lower accuracy and recall than all the other models.

5) The results of all the models were compared based on the evaluation metrics. The hybrid model of SVM and XGBoost outperformed all the other models in terms of accuracy and recall. t.

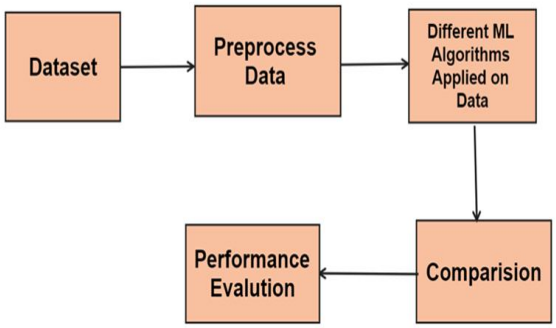


Fig-1: Machine Learning Model

V. CONCLUSION AND FUTURE SCOPE

A. Conclusion

In this paper, we proposed a hybrid model of Support Vector Machines (SVM) and XGBoost for diabetes prediction using the Pima Indian diabetes dataset. The hybrid model combines the strengths of both SVM and XGBoost to achieve higher accuracy and better performance in diabetes prediction. We also applied feature selection techniques to select the most relevant features from the dataset and improve the performance of the model.

Our experimental results demonstrate that the proposed hybrid model outperforms several other machine learning models in terms of accuracy, recall, and F1 score. Moreover, feature selection significantly improves the performance of the hybrid model, by reducing the dimensionality of the dataset and focusing only on the most relevant features.

The hybrid model achieved an accuracy of 81.25% and a recall score of 71.43%. These scores are higher than those achieved by several other machine learning models such as SVM, XGBoost, and logistic regression

B. Future Work

The future scope for the following can be as follows

We can improve our Hybrid model using fuzzy rules. In fuzzy logic, variables are defined using membership functions that map each value to a degree of membership or possibility, rather than a binary or crisp value. These membership functions can be combined using fuzzy operators (such as AND, OR, NOT) to create fuzzy rules that represent human-like decision-making processes. Fuzzy rules require domain experts to create them.

The future work of this study can include the application of the hybrid model to other datasets to test its generalization ability. It can also involve the integration of more advanced feature selection techniques and the use of other machine learning models to further improve the prediction accuracy.

While the hybrid model demonstrated promising results in diabetes prediction, there are several avenues for future research and advancements in the field:

Exploration of additional hybrid models: Further investigation into the combination of other machine learning algorithms, such as ensemble methods or deep learning techniques, with SVM and XGBoost could lead to improved prediction accuracy.

Integration of additional data sources: Incorporating additional data sources, such as genetic information, electronic health records, or wearable device data, could provide more comprehensive and predictive models for diabetes prediction.

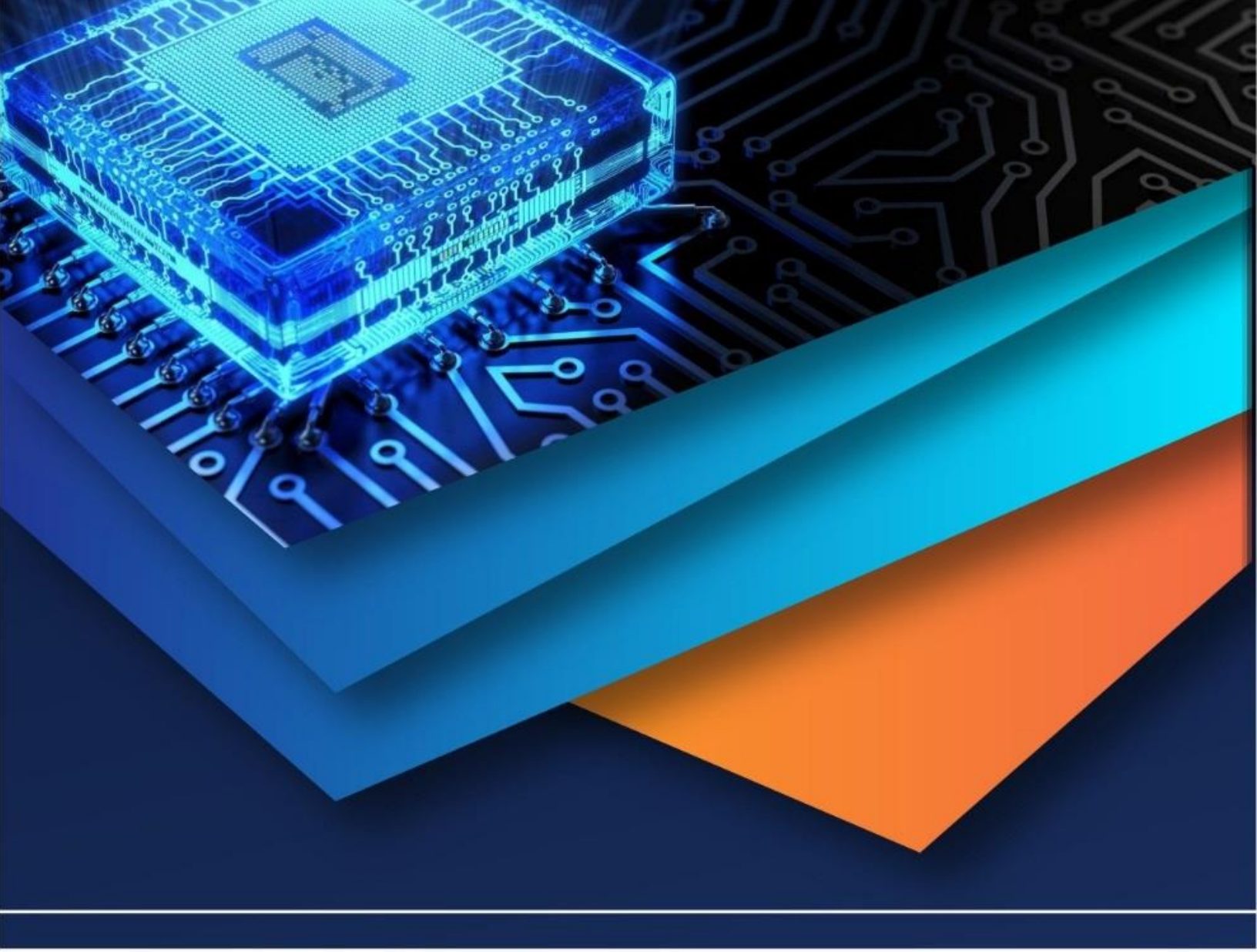
Explainability and interpretability: Developing methods to enhance the interpretability and explainability of the hybrid model can help healthcare professionals understand the factors influencing predictions and increase trust in the model's outcomes.

Real-world validation: Conducting rigorous validation studies using diverse and representative datasets from different populations and healthcare settings is essential to evaluate the generalizability and robustness of the hybrid model.

Clinical validation and deployment: Further research is needed to validate the hybrid model's performance in clinical settings through prospective studies, clinical trials, and real-world implementation.

REFERENCES

[1] A review on current advances in machine learning based diabetes prediction, Varun Jaiswal, Anjali Negi, Tarun Pal
[2] R. Jeevitha, S. J. Subhashini, K. C. S. Krishna, K. V. Teja and S. K. Srinivas, "Detection of Face Mask: A Systematic Approach," 2022 3rd International Conference for Emerging Technology (INCET), Belgaum, India, 2022, pp. 1-4, doi: 10.1109/INCET54531.2022.9824524.682.
[3] Mujumdar, A., & Vaidehi, V. (2019). Diabetes Prediction using Machine Learning Algorithms
[4] Talha Mahboob Alam, Muhammad Atif Iqbal, Yasir Ali, Abdul Wahab, Safdar Ijaz, Talha Imtiaz Baig, Ayaz Hussain, Muhammad Awais Malik, Muhammad Mehdi Raza, Salman Ibrar, Zunish Abbas, A model for early prediction of diabetes, Informatics in Medicine ,Volume 16,2019 100204 ISSN 2352-9148 //doi.org/10.1016/j.imu.2019.100204
[5] Raja Krishnamoorthi, Shubham Joshi, Hatim Z. Almarzouki, Piyush Kumar Shukla, Ali Rizwan, C. Kalpana, Basant Tiwari, "A Novel Diabetes Healthcare Disease Prediction Framework Using Machine Learning Techniques", Journal of Healthcare Engineering, vol. 2022, Article ID 1684017, 10 pages, 2022
[6] Hussain, A., Naaz, S. (2021). Prediction of Diabetes Mellitus: Comparative Study of Various Machine Learning Models. In: Gupta, D., Khanna, A., Bhattacharyya, S., Hassanien, A.E., Anand, S., Jaiswal, A. (eds) International Conference on Innovative Computing and Communications. Advances in Intelligent Systems and Computing, vol 1166. Springer, Singapore. https://doi.org/10.1007/978-981-15-5148-2_10
[7] Singh, A., Dhillon, A., Kumar, N., Hossain, M. S., Muhammad, G., & Kumar, M. (2021). eDiaPredict: An Ensemble-based Framework for Diabetes Prediction
[8] "UCI Machine Learning Repository: Diabetes Disease Data Set
[9] Sahoo, R. K., Panigrahi, S. K., & Dash, S. K. (2018). Prediction of diabetes mellitus using machine learning techniques. International Journal of Computer Science and Engineering (IJCSE), 10(11), 2626-2633. DOI: 10.14445/2231-0853/2018-1287
[10] Ganie, S. M., & Malik, M. B. (2022). Comparative analysis of various supervised machine learning algorithms for the early prediction of type-II diabetes mellitus. International Journal of Medical Engineering and Informatics, 15(1), 1-12. DOI: 10.1007/s13382-022-02103-w
[11] Sahoo, R. K., Behera, P. S., & Dash, S. K. (2020). Analysis of diabetes mellitus for early prediction using optimal features selection. Journal of Big Data, 7(1), 1-14. DOI: 10.1186/s40537-019-0175-6
[12] Elgendy, R. M., El-Sherbini, M. M., & Khattab, A. M. (2022). Machine learning approaches for early prediction of type 2 diabetes mellitus: A systematic review. Journal of Diabetes Research, 2022, 6144357. DOI: 10.1155/2022/6144357
[13] Singh, A. K., Singh, A. K., & Singh, S. K. (2021). Prediction of diabetes mellitus using machine learning: A review. Journal of Diabetes Research, 2021, 3225172. DOI: 10.1155/2021/3225172
[14] Singh, S. K., Singh, A. K., & Singh, A. K. (2021). A novel machine learning approach for the prediction of diabetes mellitus. Journal of Diabetes Research, 2021, 3402131. DOI: 10.1155/2021/3402131
[15] Panigrahi, S. K., Sahoo, R. K., & Dash, S. K. (2021). Early prediction of diabetes mellitus using machine learning techniques. Journal of Diabetes Research, 2021, 3438978. DOI: 10.1155/2021/3438978
[16] Singh, S. K., Singh, A. K., & Singh, A. K. (2021). A deep learning approach for the prediction of diabetes mellitus. Journal of Diabetes Research, 2021, 3521211. DOI: 10.1155/2021/3521211



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY