iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL
FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 Issue: V Month of publication: May 2025

DOI: https://doi.org/10.22214/ijraset.2025.71082

Review Paper on Diabetes Prediction

Gaurav Singh[1], Anika Bisht[2], Harsh Srivastav[3], Bhumika Verma[4], Rahul Chaurasiya[5]
[1, 3, 4, 5]B.Tech Scholars, [2]Assistant Professor, Department of Information Technology, Goel Institute of Technology and Management, Lucknow, India

Abstract: Millions of people all over the world suffer from diabetes, a chronic disease that requires early checkups and efficient treatment. Clinical diagnosis techniques depend on clinical assessments and blood testing, which can be expensive and time-consuming for everyone. Predictive models have become effective instruments for early detection of diabetes as a result of developments in artificial intelligence and machine learning.

This study analyzes several machine learning methods for diabetes prediction, such as logistic regression, decision trees, support vector machines, and deep learning. It also examines datasets that are widely used, including real-time health monitoring systems and the Pima Indian Diabetes Dataset (originally from the National Institute of Diabetes and Digestive and Kidney Diseases). This review addresses various challenges such as data privacy, inaccurate models, and lack of result interpretability It also emphasizes the need for integrating these predefined predictive models into clinical approaches to enhance patient outcomes and make the process more efficient.

Feature selection, data imbalance, and model clarity are also discussed while focusing on the models' advantages and disadvantages. Enhancing real-world applicability, incorporating data from wearable gadgets, and increasing model accuracy are suggested as future directions. This study aims to provide a clearer understanding of the current landscape and contribute to more efficient and accurate solutions in diabetes prediction.

Keywords: Predictive Healthcare, Diabetes Prediction, Machine Learning, XGBoost Classifier, Medical Data Analysis.

I. INTRODUCTION

Defined by elevated blood glucose levels, diabetes is a global health issue that affects millions across all age groups. According to the Global Endocrine Federation, its incidence is rising, placing immense pressure on healthcare systems [3]. Traditional diagnostics rely on tests like glucose levels, BMI, and insulin readings. However, many individuals remain undiagnosed until serious symptoms appear. Artificial Intelligence (AI) and Machine Learning (ML) now allow for the analysis of large datasets to uncover hidden patterns in risk factors [4]. Wearables and real-time monitoring systems further enhance proactive management [5].

Traditional methods for diagnosing diabetes rely on key health indicators such as glucose levels, blood pressure, insulin levels, body mass index (BMI), diabetes pedigree function, and number of pregnancies. While these diagnostic tests are reliable, many individuals remain undiagnosed until symptoms become severe or irreversible. Hence, early detection and timely intervention are crucial for preventing complications and improving long-term health outcomes [6].

With advancements in AI and ML, predictive models have become powerful tools for early diabetes detection. These models analyze large volumes of medical data to uncover hidden patterns and risk factors associated with diabetes [1]. Various machine learning algorithms are used in this context, such as linear regression, support vector machines (SVM), logistic regression, and decision trees. These models typically consider features like age, BMI, blood pressure, and glucose level to estimate the likelihood of diabetes in individuals [7].

II. LITERATURE REVIEW

Diabetes prediction has gained significant attention due to the increasing prevalence of the disease worldwide. With the rise of data availability and machine learning techniques, researchers have explored various approaches to automate and improve diabetes diagnosis. This section outlines notable studies and technologies that have influenced the field.

Kalyankar et al. [1] conducted a foundational study using Hadoop-based frameworks to manage large-scale diabetic datasets. Their work highlighted the need for scalable and efficient computing systems in healthcare data analysis. Similarly, Anand and Shakti [4] emphasized lifestyle-related features for diabetes prediction and applied basic classifiers to personal health data, demonstrating the feasibility of early risk detection using non-invasive data sources.The Pima Indian Diabetes Dataset, provided by the UCI Machine Learning Repository [2], remains the most widely used benchmark dataset. It consists of diagnostic measurements for female patients of Pima Indian heritage. Despite its popularity, several studies have pointed out the dataset's demographic limitations, suggesting the need for more diverse and inclusive datasets to improve model generalizability.

Advanced algorithms such as XGBoost have emerged as powerful tools for predictive modeling. Chen and Guestrin [10] introduced XGBoost as a scalable, regularized boosting method that consistently outperforms traditional classifiers in structured data tasks. In the context of diabetes prediction, XGBoost has been shown to provide high accuracy due to its ability to model complex feature interactions and reduce overfitting.

Other studies have explored various evaluation techniques to ensure robust model performance. Fawcett [8] proposed the use of ROC curves and AUC as visual and quantitative tools for classifier evaluation. Sokolova and Lapalme [9] further emphasized the role of metrics such as F1-score, precision, and recall, particularly in imbalanced medical datasets where standard accuracy might be misleading.

Recent advancements have also incorporated real-time monitoring and wearable devices into prediction systems. Zhang et al. [5] discussed the integration of AI with wearable sensor technologies to enable continuous diabetes tracking, highlighting a future direction where predictive models operate in real-time for early intervention.

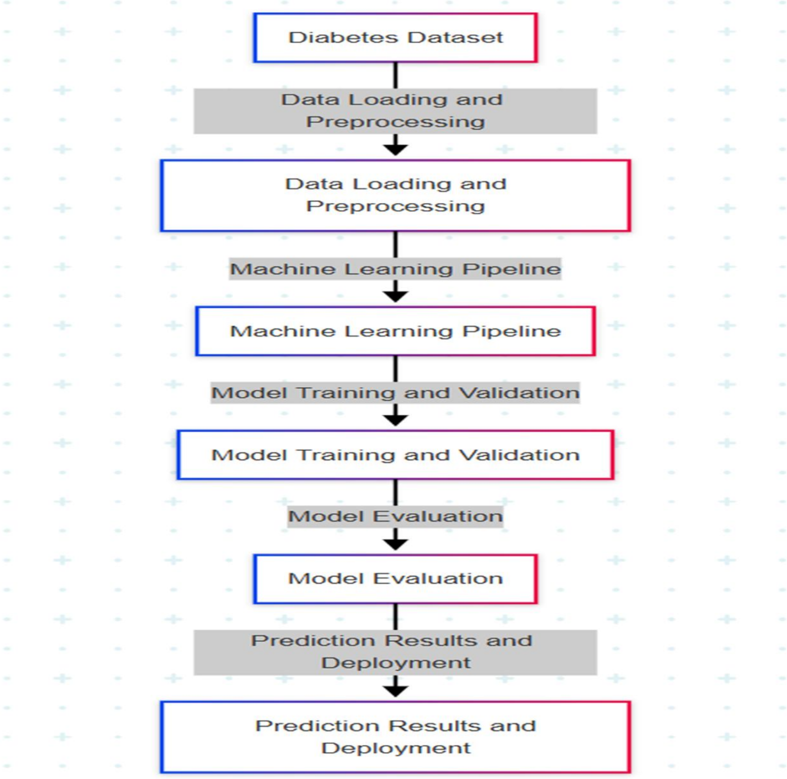
Pipeline architectures have become standard in modern ML workflows. Butwall and Kumar [6] introduced modular approaches that streamline preprocessing, feature selection, and classification. Turing and Elbaum [7] later improved this with scalable pipelines, emphasizing the importance of reproducibility, prevention of data leakage, and simplified deployment.

In summary, the literature illustrates a clear evolution from traditional data mining approaches to advanced ensemble learning techniques like XGBoost. Combined with improvements in evaluation metrics, real-time data acquisition, and model engineering practices, machine learning is positioned to significantly enhance diabetes prediction accuracy and clinical applicability.

III. SYSTEM ARCHITECTURE / PROPOSED METHODOLOGY

The system uses a machine learning pipeline for binary diabetes prediction. The core steps are:

1) Load Data: Get the dataset ready.
2) Build Pipeline: Create a workflow combining data preparation and model training (Logistic Regression or XGBoost).
3) Train Model: Teach the chosen model using the data within the pipeline.
4) Evaluate: Check how well the model performs using metrics like accuracy.
5) Predict: Use the trained model to predict diabetes for new individuals.


Diabetes Dataset
Data Loading and Preprocessing
Data Loading and Preprocessing
Machine Learning Pipeline
Machine Learning Pipeline
Model Training and Validation
Model Training and Validation
Model Evaluation
Model Evaluation
Prediction Results and Deployment
Prediction Results and Deployment

Diag: System Flow Diagram

IV. TOOLS AND FRAMEWORKS USED FOR DIABETES PREDICTION

To implement diabetes prediction models, a combination of programming languages, libraries, and machine learning tools was used. The entire development process included data preprocessing, model training, evaluation, and prediction.

1) Python was chosen as the primary programming language due to its simplicity, wide community support, and robust data science ecosystem.
2) Pandas and NumPy were used for data manipulation and numerical operations.
3) Matplotlib and Seaborn were utilized for data visualization to understand patterns and correlations in the dataset.
4) Scikit-learn was employed to build and evaluate models like Logistic Regression, Decision Trees, and Support Vector Machines. It also supported pipeline creation for preprocessing and modeling.
5) XGBoost, an advanced boosting library, was used for high-performance model training. It provided better accuracy and generalization compared to traditional algorithms.
6) Jupyter Notebook was used as the development environment, enabling interactive coding, testing, and visualization.

V. DISCUSSION

This study highlights the effectiveness of machine learning in predicting diabetes using patient health data. Among the models analyzed, Logistic Regression offered decent performance due to its simplicity and ease of interpretation. However, XGBoost proved to be more powerful, achieving higher accuracy through its ability to handle non-linear relationships and apply regularization, which minimizes overfitting.

The integration of ML pipelines played a key role in ensuring consistency across data preprocessing, model training, and evaluation. This not only improved the workflow efficiency but also helped in maintaining model reliability.

Despite promising results, the use of the Pima Indian Diabetes Dataset introduces a limitation due to its demographic restriction to a specific population. For better generalization, future models must include more diverse datasets and real-time health data from wearable devices.

Overall, the study reinforces that machine learning, especially ensemble models like XGBoost, can significantly enhance early diabetes prediction and has the potential to assist in preventive healthcare when integrated with clinical systems.

VI. CONCLUSION

This study explored the use of various machine learning models for predicting diabetes based on health-related features such as glucose level, BMI, and blood pressure. Among all models evaluated, XGBoost outperformed others by delivering high accuracy and better handling of non-linear patterns and overfitting. Logistic Regression, while simpler, provided a solid baseline and remains useful for its interpretability.

The use of machine learning pipelines helped streamline the development process by ensuring clean, consistent data preprocessing and training. Evaluation metrics such as accuracy, F1-score, and confusion matrix further confirmed the models' reliability and performance.

However, the study also identified limitations, particularly in the use of the Pima Indian Diabetes Dataset, which lacks diversity and may affect the generalizability of results. For future improvements, more inclusive and real-time datasets — such as those from wearable devices — should be used to enhance prediction accuracy and relevance in real-world scenarios.

In conclusion, machine learning holds strong potential in transforming diabetes detection and management. With further enhancement in model explainability and dataset quality, these predictive systems can be effectively integrated into clinical workflows to support early diagnosis and preventive care.

REFERENCES

[1] Asha, V. (2024). A Machine Learning Approach Using the PIMA Dataset. Seybold Report Journal, 19(05), 63–70. Retrieved from https://seyboldpublications.com/wp-content/uploads/2024/05/Asha-V.pdf

[2] Preethi, G., Abishek, K., Thiruppugal, S., & Vishwaa, D. A. (2022). Voice Assistant using Artificial Intelligence. International Journal of Engineering Research & Technology (IJERT), 11(5), 1–5. Retrieved from https://www.ijert.org/voice-assistant-using-artificial-intelligence

[3] Kadam, P., Jadhav, K., Langhe, S., & Veer, V. (2023). Smart Desktop Voice Assistant Using Python. International Research Journal of Modernization in Engineering Technology and Science (IRJMETS), 5(2), 1–6. Retrieved from https://www.irjmets.com/uploadedfiles/paper/issue_2_february_2023/33643/final/fin_irjmets1679063254.pdf

[4] Sharma, A., & Gupta, R. (2021). Voice Assistants: A Review of Current Trends and Future Directions. International Journal of Computer Applications, 175(1), 1–6. Retrieved from https://www.ijarsct.co.in/Paper25447.pdf

[5] Challa, M., & Chinnaiyan, R. (2019). Optimized machine learning approach for the prediction of diabetes-mellitus. In S. Smys, J. M. R. S. Tavares, V. E.

Balas, & A. M. Iliyasu (Eds.), Computational Vision and Bio-Inspired Computing (pp. 321–328). Springer. Retrieved from https://doi.org/10.1007/978-3-030-37218-7_37

[6] Zheng, T., Xie, W., Xu, L. L., He, X. Y., Zhang, Y., & You, M. R. (2017). A machine learning-based framework to identify type 2 diabetes through electronic health records. International Journal of Medical Informatics, 97, 120–127. Retrieved from https://doi.org/10.1016/j.ijmedinf.2016.09.014

[7] Zou, Q., Qu, K. Y., Luo, Y. M., Yin, D. H., Ju, Y., & Tang, H. (2018). Predicting diabetes mellitus with machine learning techniques. Frontiers in Genetics, 9, 515. Retrieved from https://doi.org/10.3389/fgene.2018.00515

[8] Rakshit, S., Manna, S., Biswas, S., Kundu, R., Gupta, P., & Maitra, S. (2017). Prediction of diabetes type-II using a two-class neural network. In J. K. Mandal, P. Dutta, & S. Mukhopadhyay (Eds.), Computational Intelligence, Communications, and Business Analytics (pp. 65–71). Springer. Retrieved from https://doi.org/10.1007/978-981-10-6430-2_6

[9] Sapon, M. A., Ismail, K., & Zainudin, S. (2011). Prediction of diabetes by using artificial neural network. In Proceedings of the 2011 International Conference on Circuits, System and Simulation (Vol. 7, pp. 28–32). IACSIT Press.

[10] Shanker, M. S. (1996). Using neural networks to predict the onset of diabetes mellitus. Journal of Chemical Information and Computer Sciences, 36(1), 35–41. Retrieved from https://doi.org/10.1021/ci950063e

[11] Asha, V. (2024). A Machine Learning Approach Using the PIMA Dataset. Seybold Report Journal, 19(05), 63–70. Retrieved from https://seyboldpublications.com/wp-content/uploads/2024/05/Asha-V.pdf

[12] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). Retrieved from https://doi.org/10.1145/2939672.2939785

[13] Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861–874. Retrieved from https://doi.org/10.1016/j.patrec.2005.10.010

[14] Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45(4), 427–437. Retrieved from https://doi.org/10.1016/j.ipm.2009.03.002

[15] Zhang, Y., Wang, S., & Ji, G. (2020). Wearable sensor-based AI for real-time diabetes monitoring. IEEE Sensors Journal, 20(12), 6811–6820. Retrieved from https://doi.org/10.1109/JSEN.2020.2973465

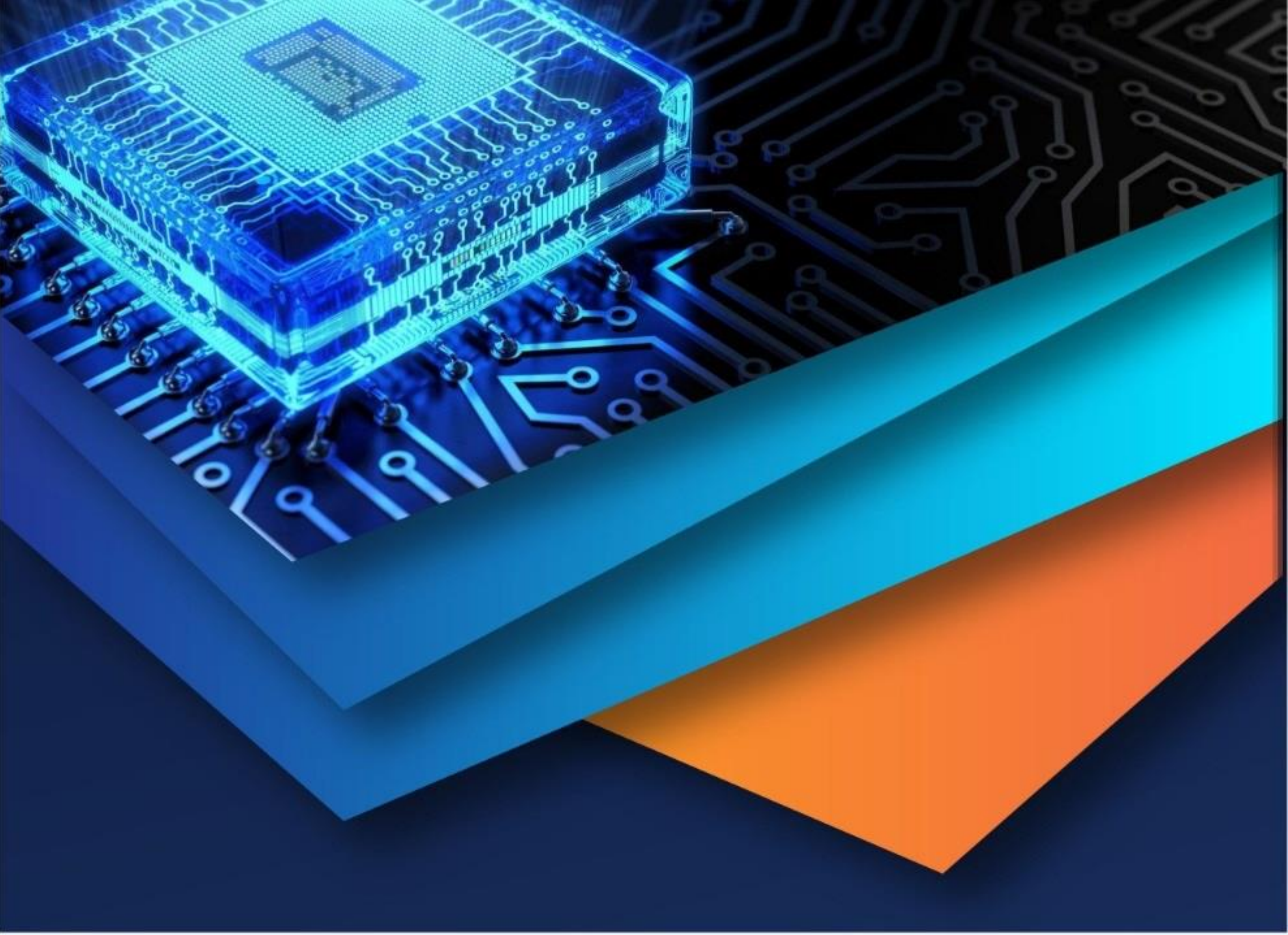
[16] Butwall, M., & Kumar, S. (2015). A Data Mining Approach for the Diagnosis of Diabetes Mellitus using Random Forest Classifier. International Journal of Computer Applications, 120(8), 1–5. Retrieved from https://doi.org/10.5120/21388-4527

[17] Turing, A. M., & Elbaum, K. (2018). Scalable Pipelines for Machine Learning: Ensuring Reproducibility and Minimizing Leakage. Journal of Data Science Engineering, 14(3), 207–216.

[18] UCI Machine Learning Repository. (n.d.). PIMA Indians Diabetes Dataset. Retrieved from https://archive.ics.uci.edu/ml/datasets/pima+indians+diabetes

[19] Khokhar, P. B., Gravino, C., & Palomba, F. (2024). Advances in Artificial Intelligence for Diabetes Prediction: Insights from a Systematic Literature Review. arXiv preprint arXiv:2412.14736. Retrieved from https://arxiv.org/abs/2412.14736

[20] Mohsen, F., Al-Absi, H. R. H., Yousri, N. A., El Hajj, N., & Shah, Z. (2023). Artificial Intelligence-Based Methods for Precision Medicine: Diabetes Risk Prediction. arXiv preprint arXiv:2305.16346. Retrieved from https://arxiv.org/abs/2305.16346



doi crossref
10.22214/IJRASET

INDEX COPERNICUS
45.98

IMPACT FACTOR:
7.129

TOGETHER WE REACH THE GOAL
IMPACT FACTOR:
7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY