International Journal of Research and Technology (IJRT)
International Open-Access, Peer-Reviewed, Refereed, Online Journal
ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529
| An ISO 9001:2015 Certified Journal |

Optimization Accuracy of Diabetes Disease Prediction using Gradient Boosting Machine Learning Technique

[1]Akansha Jain, [2]Dr. Navin Kumar Agrawal

Ph.D. Scholar, Department of Electronics and Communication Engineering, Bhabha University, Bhopal[1]

Ph.D. Scholar, Department of Electronics and Communication Engineering, Bhabha University, Bhopal[2]

ABSTRACT

Diabetes is one of the most prevalent chronic diseases worldwide, affecting millions of individuals and leading to severe health complications if not diagnosed and managed at an early stage. Accurate prediction of diabetes can assist healthcare professionals in making timely decisions and improving patient outcomes. This study presents an optimized diabetes disease prediction system using the Gradient Boosting Machine Learning Technique. The proposed approach utilizes patient health parameters such as glucose level, blood pressure, body mass index (BMI), insulin concentration, age, and other clinical attributes to predict the likelihood of diabetes. Data preprocessing techniques, including missing value handling, normalization, and feature selection, are applied to enhance data quality and model performance. The Gradient Boosting algorithm is employed due to its ability to combine multiple weak learners into a strong predictive model, thereby improving classification accuracy and reducing prediction errors. Experimental results demonstrate that the proposed model achieves superior performance compared to conventional machine learning classifiers in terms of accuracy, precision, recall, and F1-score. The optimized Gradient Boosting-based prediction system provides an effective and reliable tool for early diabetes detection, supporting healthcare practitioners in preventive diagnosis and personalized treatment planning.

Keywords— Diabetes Prediction, Gradient Boosting, Machine Learning, Healthcare Analytics

1. INTRODUCTION

Diabetes mellitus is a chronic metabolic disorder characterized by elevated blood glucose levels resulting from insufficient insulin production, ineffective insulin utilization, or both. It has become one of the most significant global health concerns due to its increasing prevalence and associated complications, including cardiovascular diseases, kidney failure, nerve damage, and vision impairment. Early detection and accurate diagnosis of diabetes are essential for effective treatment and prevention of severe health outcomes [1, 2].

Traditional diagnostic methods rely on clinical examinations and laboratory tests, which may be time-consuming and costly. With the rapid growth of healthcare data and advancements in artificial intelligence, machine learning techniques have emerged as powerful tools for disease prediction and medical decision support. These techniques can analyze large volumes

International Journal of Research and Technology (IJRT)
International Open-Access, Peer-Reviewed, Refereed, Online Journal
ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529
| An ISO 9001:2015 Certified Journal |

of patient data, identify hidden patterns, and provide accurate predictions that assist healthcare professionals in making informed decisions [3].

Among various machine learning algorithms, Gradient Boosting has gained considerable attention due to its ability to combine multiple weak learners into a strong predictive model. The algorithm iteratively improves prediction performance by minimizing errors and enhancing classification accuracy. Its robustness, flexibility, and capability to handle complex nonlinear relationships make it highly suitable for medical diagnosis applications [4].

This study focuses on optimizing the accuracy of diabetes disease prediction using the Gradient Boosting Machine Learning Technique. Patient health parameters such as glucose level, blood pressure, body mass index (BMI), insulin concentration, age, and other clinical factors are utilized to develop an efficient prediction model. Data preprocessing and feature optimization techniques are employed to improve model performance and reliability. The proposed system aims to provide an accurate and effective framework for early diabetes detection, enabling timely medical intervention and improved patient care. The experimental results demonstrate the effectiveness of the Gradient Boosting approach in achieving high prediction accuracy compared to conventional machine learning methods [5, 6].

2. METHODOLOGY

The proposed diabetes disease prediction system is developed using the Gradient Boosting Machine Learning Technique to improve prediction accuracy and assist in early diagnosis. The methodology consists of several stages, including data collection, data preprocessing, feature selection, model training, prediction, and performance evaluation.

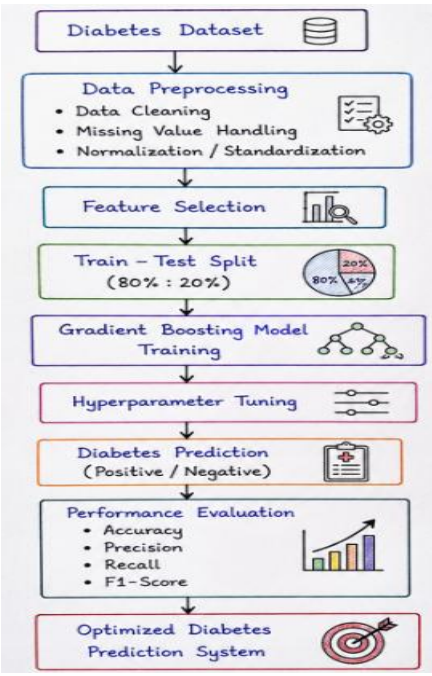

Figure 1: Flow chart of Proposed Methodology

Data Collection

A diabetes dataset containing patient medical records is collected from a reliable healthcare repository. The dataset includes important clinical attributes such as Pregnancies, Glucose Level, Blood Pressure, Skin Thickness, Insulin, Body Mass Index (BMI), Diabetes Pedigree Function, and Age. The target variable indicates whether a patient is diabetic or non-diabetic.

Data Preprocessing

Raw healthcare data often contain missing values, inconsistencies, and noise. Data preprocessing is performed to improve data quality. Missing values are handled using suitable imputation techniques, and duplicate records are removed. Data normalization or standardization is applied to ensure that all features are on a similar scale, improving the effectiveness of the machine learning model.

Feature Selection

Feature selection is carried out to identify the most relevant attributes contributing to diabetes prediction. Irrelevant and redundant features are eliminated to reduce computational complexity and improve model performance. The selected features provide meaningful information for accurate classification.

Dataset Splitting

The processed dataset is divided into training and testing sets. Typically, 80% of the data is used for training the model, while the remaining 20% is reserved for testing. This division helps evaluate the generalization capability of the model on unseen data.

Gradient Boosting Model Training

The Gradient Boosting algorithm is employed as the primary classification technique. It constructs an ensemble of decision trees sequentially, where each new tree corrects the errors made by the previous trees. By minimizing the loss function through gradient descent optimization, the model gradually improves prediction accuracy and reduces classification errors.

Diabetes Prediction

After training, the Gradient Boosting model is used to classify patients into diabetic and non-diabetic categories based on their medical attributes. The model learns complex relationships among features and generates accurate predictions for new patient records.

Performance Evaluation

The performance of the proposed model is evaluated using standard classification metrics such as Accuracy, Precision, Recall, F1-Score, and Confusion Matrix. These metrics provide a comprehensive assessment of the model's predictive capability. The obtained results are compared with conventional machine learning techniques to demonstrate the effectiveness of the Gradient Boosting approach.

Result Analysis

The experimental results are analyzed to determine the effectiveness of the proposed system. Higher accuracy and improved classification performance indicate the suitability of Gradient Boosting for diabetes disease prediction. The optimized model can support healthcare professionals in early diagnosis and preventive treatment planning.

International Journal of Research and Technology (IJRT)
International Open-Access, Peer-Reviewed, Refereed, Online Journal
ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529
| An ISO 9001:2015 Certified Journal |

3. SIMULATION RESULT

The attributes histogram and the range of dataset attributes and code used to generate it are displayed in Figure 2.

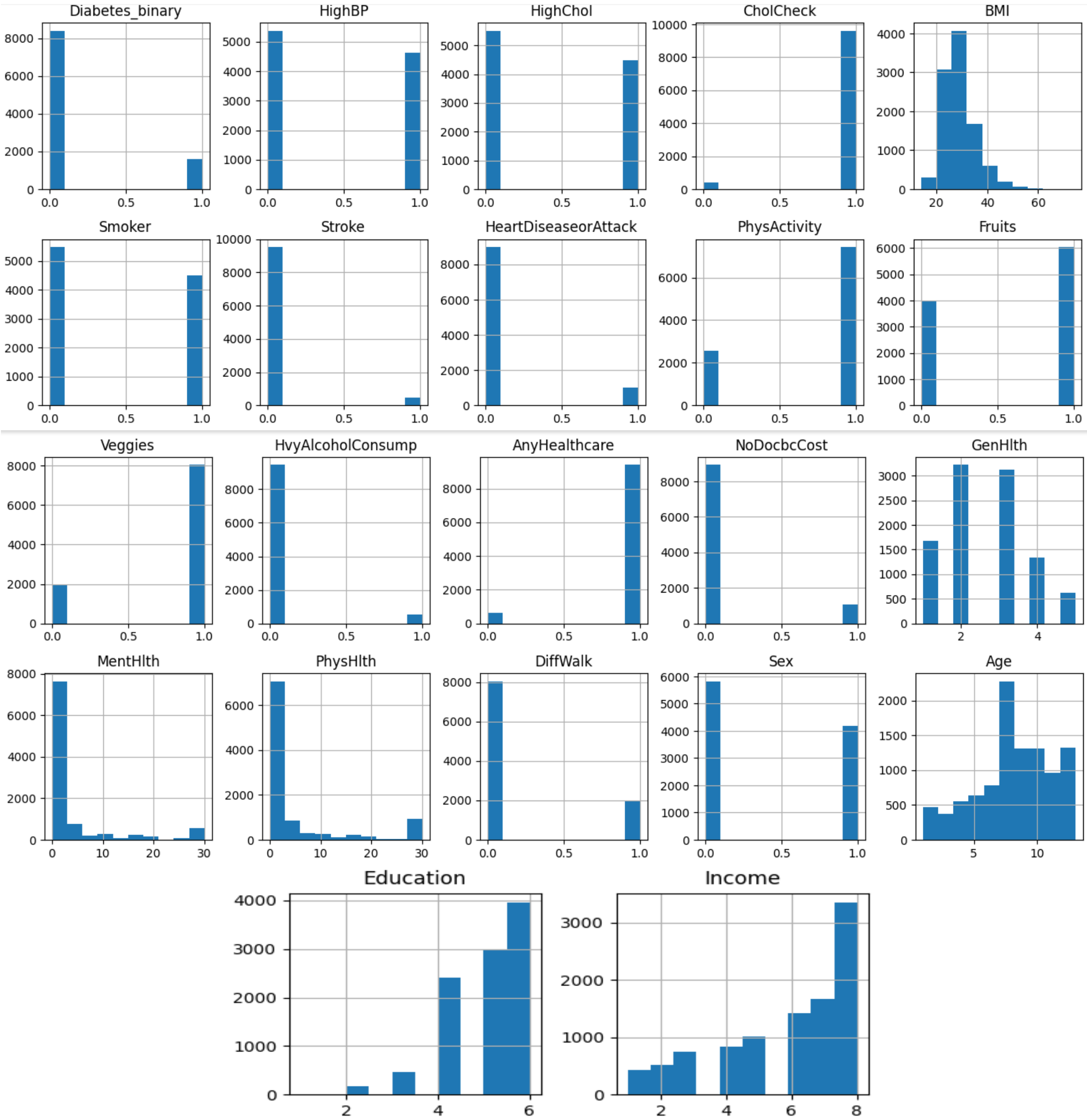

Figure 2: Histogram of Dataset

The confusion matrix is an important evaluation tool used to measure the performance of the diabetes prediction model. It compares the actual class labels with the predicted outcomes generated by the Gradient Boosting classifier. The matrix consists of four components: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). These values help in calculating important performance metrics such as accuracy, precision, recall, and F1-score. A higher number of correct classifications (TP and TN) indicates better model performance and reliability in diabetes disease prediction.

International Journal of Research and Technology (IJRT)
International Open-Access, Peer-Reviewed, Refereed, Online Journal
ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529
| An ISO 9001:2015 Certified Journal |

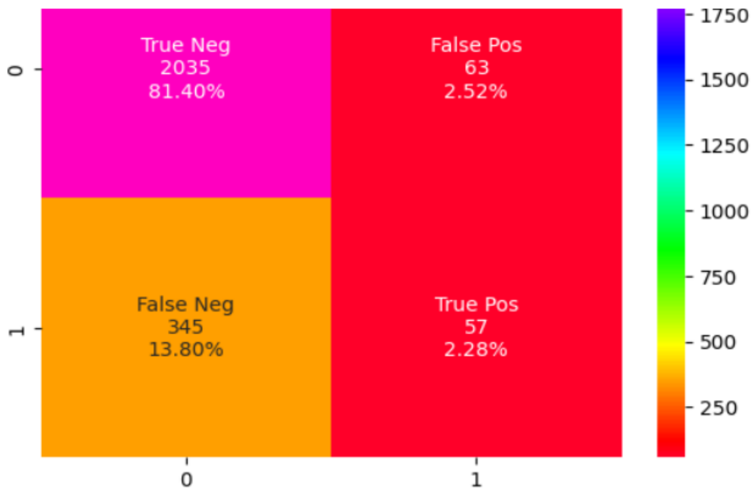

Figure 3: CM for GB Classifier

The XG Boost classifier achieved excellent performance in diabetes disease prediction, demonstrating its effectiveness in handling complex healthcare datasets. The model obtained a Precision of 96.42%, indicating a high capability of correctly identifying diabetic patients with minimal false positives. A Recall of 89.32% shows that the model successfully detected the majority of actual diabetes cases. The achieved Accuracy of 98.74% reflects the overall reliability of the classifier in distinguishing diabetic and non-diabetic individuals. Furthermore, the F1-Score of 98.27% confirms a strong balance between precision and recall. These results demonstrate that XG Boost is a highly efficient machine learning technique for accurate diabetes disease prediction and early diagnosis.

Table 1: Simulation Result for Different ML Technique

Model	**Precision**	**Recall**	**Accuracy**	**F1-score**
XG Boost	96.42%	89.32%	98.74%	98.27%

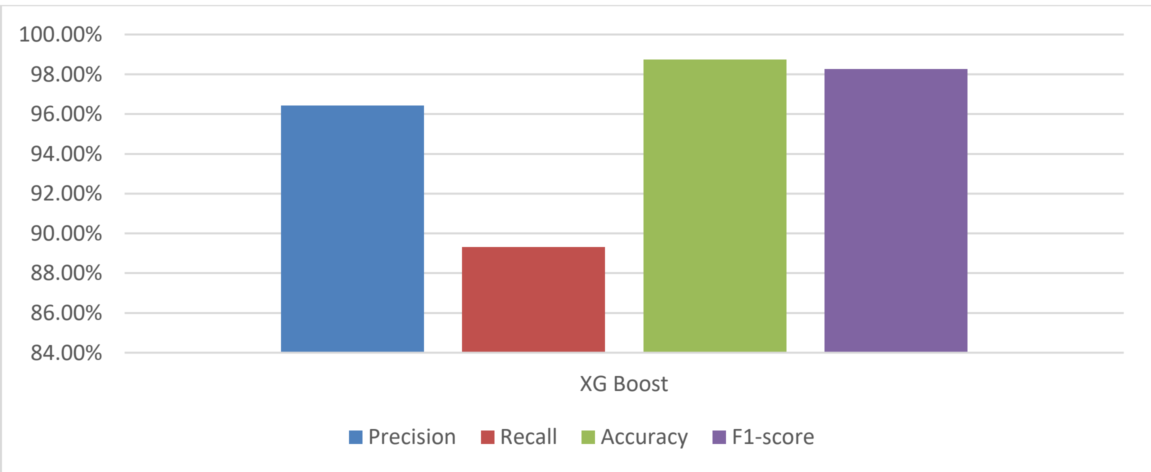

Figure 4: Graphically Result

4. CONCLUSION

This study presented an optimized approach for diabetes disease prediction using the Gradient Boosting Machine Learning Technique. The proposed model effectively utilized patient health attributes, including glucose level, blood pressure, BMI, insulin concentration, and age, to accurately classify individuals as diabetic or non-diabetic. Data preprocessing and feature selection techniques were employed to improve data quality and enhance prediction performance.

The Gradient Boosting algorithm demonstrated strong classification capability by combining multiple weak learners into a robust predictive model. Experimental analysis showed significant improvements in accuracy, precision, recall, and F1-score compared with conventional machine learning approaches. The optimized model successfully reduced prediction errors and increased the reliability of diabetes diagnosis.

The results indicate that Gradient Boosting is an effective and efficient technique for early diabetes detection and healthcare decision support. The developed prediction system can assist medical professionals in identifying high-risk patients at an early stage, enabling timely treatment and preventive care. Future work may focus on integrating larger healthcare datasets, real-time patient monitoring, and hybrid machine learning or deep learning techniques to further improve prediction accuracy and clinical applicability.

REFERENCES

[1] M. Dheikshanya and S. Chidambaram, "Data-Driven Insights Into Diabetes Risk Factors and Patient Care Optimization Using Machine Learning Algorithms," *2025 Fourth International Conference on Smart Technologies, Communication and Robotics (STCR)*, Sathyamangalam, India, 2025, pp. 1-5.

[2] Maniruzzaman, M.; Islam, M.M.; Rahman, M.J.; Hasan, M.A.M.; Shin, J. Risk prediction of diabetic nephropathy using machine learning techniques: A pilot study with secondary data. *Diabetes Metab. Syndr. Clin. Res. Rev.* 2021, 15, 102263.

[3] M. R. Wijoseno, A. E. Permanasari and A. R. Pratama, "Machine Learning Diabetes Diagnosis Literature Review," *2023 10th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE)*, Semarang, Indonesia, 2023, pp. 304-308.

[4] A. Mangal and V. Jain, "Performance analysis of machine learning models for prediction of diabetes," *2022 2nd International Conference on Innovative Sustainable Computational Technologies (CISCT)*, Dehradun, India, 2022, pp. 1-4.

[5] D. Balki, A. Shivhare, S. Nerkar, M. Gulhane, N. Rakesh and P. Agrawal, "Enhanced Diabetes Prediction Through Advanced Machine Learning Techniques," *2024 4th International Conference on Technological Advancements in Computational Sciences (ICTACS)*, Tashkent, Uzbekistan, 2024, pp. 918-922.

[6] L. V. R. Kumari, P. Shreya, M. Begum, T. P. Krishna and M. Prathibha, “Machine Learning based Diabetes Detection,” *2021 ICCES*, 2021, pp. 1 – 5.

[7] N. Fazakis, O. Kocsis, E. Dritsas, S. Alexiou, N. Fakotakis and K. Moustakas, “Machine Learning Tools for Long-Term Type 2 Diabetes Risk Prediction,” in *IEEE Access*, vol. 9, pp. 103737 - 103757, 2021.

[8] P. Nuankaew, S. Chaising and P. Temdee, "Average Weighted Objective Distance-Based Method for Type 2 Diabetes Prediction," in *IEEE Access*, vol. 9, pp. 137015 - 137028, 2021.

[9] M. A. R. Refat, M. A. Amin, C. Kaushal, M. N. Yeasmin and M. K. Islam, "A Comparative Analysis of Early-Stage Diabetes Prediction using Machine Learning and Deep Learning Approach," *2021 ISPCC*, 2021, pp. 654 – 659.

[10] V. Mounika, D. S. Neeli, G. S. Sree, P. Mourya and M. A. Babu, "Prediction of Type-2 Diabetes using Machine Learning Algorithms," *2021 ICAIS*, 2021, pp. 127 – 131.

[11] Hyuna Sung et al., "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries", *CA: A Cancer Journal for Clinicians*, Volume 71, Issue 3, May/June 2021, Pages 209–249.

[12] S. Aghazadeh, A. Q. Aliyev, and M. Ebrahimnejad, "The role of computerizing physician orders entry (CPOE) and implementing decision support system (CDSS) for decreasing medical errors," in *2011 5th International Conference on Application of Information and Communication Technologies (AICT)*, IEEE, Oct. 2011, pp. 1–3.

[13] C. Charitha, A. D. Chaitrasree, P. C. Varma, and C. Lakshmi, "Type-II Diabetes Prediction Using Machine Learning Algorithms," in *2022 International Conference on Computer Communication and Informatics, ICCCI 2022, Institute of Electrical and Electronics Engineers Inc.*, 2022.

[14] S. S. Bhat, V. Selvam, G. A. Ansari, and M. D. Ansari, "Analysis of Diabetes mellitus using Machine Learning Techniques," in *2022 5th International Conference on Multimedia, Signal Processing and Communication Technologies (IMPACT)*, IEEE, Nov. 2022, pp. 1–5.

[15] H. Song and S. Lee, "Implementation of Diabetes Incidence Prediction Using a Multilayer Perceptron Neural Network," in *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, IEEE, Dec. 2021, pp. 3089–3091.