# MACHINE LEARNING–BASED DIABETIC CLASSIFICATION USING DATA MINING AND ENSEMBLE STRATEGIES

**K Sreedevi**
Research Scholar
Dept of Computer science and Engineering
NIILM University-Haryana
sreedevikadiyala05@gmail.com

**Dr. Gyanendra Kumar Gupta**
Professor
Department of Computer Science & Engineering
Niilm University Kaithal-Haryana
gyanendrag@gmail.com

**Prof. (Dr.) Mukesh Kumar Rana**
Professor & Head (CSE &CSA)
Department of Computer Science & Engineering
Niilm University Kaithal-Haryana
mrana91@gmail.com

## ABSTRACT

*Diabetes mellitus is a chronic health condition that poses serious health risks if not detected and managed early. Accurate and timely classification of diabetic status can significantly aid in preventive healthcare and treatment strategies. This paper proposes a robust framework integrating data mining techniques and ensemble learning methods to enhance the accuracy and reliability of diabetic classification using machine learning models. The proposed framework begins with comprehensive data pre-processing, including handling missing values, feature selection, and normalization. Key features influencing diabetic outcomes are extracted through correlation analysis and feature importance measures. Multiple machine learning models such as Decision Trees, Random Forests, Support Vector Machines (SVM), and Gradient Boosting are employed for initial classification. To further improve predictive performance, ensemble methods like Bagging, Boosting, and Voting Classifiers are implemented. The framework is evaluated using standard performance metrics such as accuracy, precision, recall, F1-score, and ROC-AUC on benchmark datasets like the Pima Indians Diabetes Dataset. Experimental results demonstrate that ensemble learning significantly outperforms individual models, offering a reliable approach for diabetic classification. This study underscores the potential of combining data mining and ensemble techniques to build intelligent and accurate diagnostic systems in healthcare.*

***Keywords:*** *Multiple machine learning models, diabetic status, diabetic classification, combining data mining and ensemble techniques*

## INTRODUCTION

Advanced computational techniques and extensive data analysis offer promising methods for classifying and forecasting diabetes outcomes. By leveraging advanced data mining techniques, Machine Learning (ML) models, and Big data technologies like Hadoop, India can improve early detection, tailor treatment plans, and manage diabetes more effectively. With the increasing availability of healthcare data and the rise of digital health systems, ML has the potential improving health outcomes for countless individuals. Diabetes mellitus (DM) has emerged as a prevalent and concerning chronic condition globally. As diabetes continues to rise in prevalence, there is a pressing need for advanced computational methods to enable timely and accurate diagnosis. Analysing healthcare data and aiding in the classification of conditions like diabetes. Traditional medical diagnostic methods frequently face limitations due to human skill and time constraints. Tasks related to categorising diabetes, resulting in varying levels of accuracy and generalisation. The intricate nature and variety of datasets associated with diabetes require robust strategies to improve prediction outcomes. Extracting valuable insights from large

databases is crucial, particularly when combined with advanced ML techniques. The identification of essential features and patterns in data, significantly boosting the effectiveness of models used for diabetes classification. Utilising data mining improves the clarity of models and the precision of predictions by leveraging relevant information and reducing complexity.

A more accurate and robust result, has become increasingly significant in the ML field over fitting, and enhance the overall effectiveness of the model. Ensemble learning offers significant benefits in classifying diabetes effectively address the unpredictability and complexity found in diabetes data. This study aims data mining techniques with ensemble learning approaches for the classification of diabetes. Classification models related to diabetes through the incorporation of advanced feature extraction, thorough data pre-processing, and ensemble learning strategies. In this context, ML has transformed the capacity to analyse healthcare data. ML approaches, particularly supervised learning, have shown efficacy in discerning patterns within extensive datasets, allowing more precise forecasts of illness outcomes. These models may be trained on annotated data, enabling computers to provide predictions and classifications based on novel, unobserved data. ML techniques may examine intricate information for diabetes categorisation, identifying trends that conventional diagnostic approaches might overlook. Recent breakthroughs in ML have markedly enhanced computers' capacity to recognise patterns, including the identification and labelling of medical pictures, voice translation, and illness outcome prediction. In diabetes management, ML can forecast a patient's probability of getting the condition, identify contributing variables, and categorise people according to their diabetes status. The capacity to automate and optimise these operations improves diagnostic accuracy and furnishes healthcare providers with essential decision-support instruments. ML has drawn a lot of interest due to its capacity to evaluate enormous volumes of medical data and spot trends that may accurately forecast the course of diseases.

## LITERATURE REVIEW

**K Ramanan et.al (2024)** Diabetes remains a critical global health concern, necessitating advanced solutions for early detection and management. This project focuses on a data mining-driven approach to develop a predictive system for diabetes using the Pima Indians Diabetes Dataset. The system incorporates a suite of data mining techniques and machine learning algorithms, including Logistic Regression, Extreme Gradient Boost (XGBoost), and Decision Tree, to analyze and interpret critical health metrics such as glucose levels, BMI, blood pressure, and insulin concentrations. The project follows a robust data mining process comprising data pre-processing, feature selection, and model evaluation. Data cleaning and normalization ensure quality and consistency, while feature selection optimizes model performance. The system applies advanced algorithms to identify hidden patterns and generate personalized diabetes risk scores, enabling early intervention and preventive care. This data mining approach empowers healthcare professionals with actionable insights through an interactive interface featuring

real-time analytics and comprehensive reporting. By leveraging data mining algorithms, this system demonstrates it's potential to enhance clinical decision-making, improve early diagnosis, and contribute to better health outcomes.

**Isfafuzzaman Tasin et.al (2023)** Many factors can cause a person to get affected by diabetes, like excessive body weight, abnormal cholesterol level, family history, physical inactivity, bad food habit etc. Increased urination is one of the most common symptoms of this disease. People with diabetes for a long time can get several complications like heart disorder, kidney disease, nerve damage, diabetic retinopathy etc. But its risk can be reduced if it is predicted early. In this paper, an automatic diabetes prediction system has been developed using a private dataset of female patients in Bangladesh and various machine learning techniques. Feature selection algorithm mutual information has been applied in this work. A semi-supervised model with extreme gradient boosting has been utilized to predict the insulin features of the private dataset. SMOTE and ADASYN approaches have been employed to manage the class imbalance problem. The authors used machine learning classification methods, that is, decision tree, SVM, Random Forest, Logistic Regression, KNN, and various ensemble techniques, to determine which algorithm produces the best prediction results.

**Aishwariya Dutta et.al (2022)** Diabetes is one of the most rapidly spreading diseases in the world, resulting in an array of significant complications, including cardiovascular disease, kidney failure, diabetic retinopathy, and neuropathy, among others, which contribute to an increase in morbidity and mortality rate. If diabetes is diagnosed at an early stage, its severity and underlying risk factors can be significantly reduced. However, there is a shortage of labeled data and the occurrence of outliers or data missingness in clinical datasets that are reliable and effective for diabetes prediction, making it a challenging endeavor. Grid search hyper parameter optimization is employed to tune the critical hyper parameters of these ML models. Furthermore, missing value imputation, feature selection, and K-fold cross-validation are included in the framework design. A statistical analysis of variance test reveals that the performance of diabetes prediction significantly improves when the proposed weighted ensemble is executed with the introduced pre-processing, with the highest accuracy of and an area under the ROC curve (AUC) of. In conjunction with the suggested ensemble model, our statistical imputation and RF-based feature selection techniques produced the best results for early diabetes prediction. Moreover, the presented new dataset will contribute to developing and implementing robust ML models for diabetes prediction utilizing population-level data.

**Umair Butt et.al (2021)** The remarkable advancements in biotechnology and public healthcare infrastructures have led to a momentous production of critical and sensitive healthcare data. By applying intelligent data analysis techniques, many interesting patterns are identified for the early and onset detection and prevention of several fatal diseases. Diabetes mellitus is an extremely life-threatening disease because it contributes to other lethal diseases, i.e., heart, kidney, and nerve damage. In this paper, a machine learning based approach has been proposed for the classification, early-stage identification,

and prediction of diabetes. Furthermore, it also presents an IoT-based hypothetical diabetes monitoring system for a healthy and affected person to monitor his blood glucose (BG) level. Moreover, a comparative analysis of the proposed approach is also performed with existing state-of-the-art techniques, demonstrating the adaptability of the proposed approach in many public healthcare applications.

**The Role Of Ml In Diabetes Prediction**

ML methods have surfaced as a possible remedy for the problems related to diabetes management and prediction. Fundamentally, ML is teaching computers to identify patterns in data so they can predict or decide without explicit programming for each job. ML has been used in the healthcare industry for a number of predictive tasks, including illness diagnosis, patient outcome prediction, and treatment plan optimisation. By examining a variety of variables, such as lifestyle decisions, medical history, and genetic predispositions, ML may assist in predicting a person's risk of acquiring diabetes. To categorise people as either low-risk or high-risk for acquiring diabetes, for instance, supervised learning algorithms like may be trained using historical data. To forecast a person's risk, these models might use a number of input factors, including family history of diabetes. One such research used an ensemble learning technique called random forests to forecast a large cohort of people's likelihood of developing diabetes based on their clinical and demographic information.

**Diabetes: A Global Health Crisis**

Despite advances in diabetes management, Type 1 diabetes remains a lifelong condition that requires intensive monitoring and management. The incidence of Type 1 diabetes has been increasing globally, though it remains less common than T2D. While the disease was historically considered a condition of older adults, its incidence has been rising among younger populations, particularly due to increasing rates of childhood obesity and sedentary lifestyles. The rising prevalence of the most significant public health concerns of our time, contributing to increasing rates of cardiovascular disease, kidney failure, and neurological disorders. Women who experience GDM are also at higher risk for complications during pregnancy and childbirth. Additionally, their children may face increased risks of obesity and metabolic disorders as they grow older.

**ML-Based Diabetes Prediction Models**

Diabetes, a chronic metabolic disorder, is characterized by elevated blood sugar levels and is a leading cause of various health complications, including cardiovascular diseases, kidney failure, and neurological issues. In recent years, ML techniques have shown great promise in predicting diabetes risk, improving early detection, and enabling timely interventions to prevent complications. Several studies have focused on employing ML methods for predicting diabetes risk by leveraging diverse datasets, including electronic health records (EHR), national surveys, and global health studies. Similarly, used data from the National Family Health Survey in India to develop a ML-based framework for predicting diabetes, highlighting the role of lifestyle factors in diabetes risk. The researchers applied support vector machines (SVM) and feature selection techniques, with results showing a significant association between lifestyle habits such as diet and physical activity and diabetes onset.

**ML for Predicting Hospital DM in Diabetic Patients**

Hospital readmission rates serve as a crucial indicator of healthcare quality, particularly for chronic conditions such as diabetes. Frequent hospital DM not only increase healthcare costs but also signal suboptimal management of the condition, potentially leading to worsened patient outcomes. Predicting DM in diabetic patients is essential for timely interventions and effective management strategies. ML algorithms, particularly ensemble and deep learning models, have emerged as effective tools in predicting DM by analyzing complex datasets derived from electronic health records (EHR) and patient history. One of the key utilized an ensemble of ML classifiers, including Random Forest (RF), Gradient Boosting, and Support Vector Machines (SVM), to predict diabetes-related hospital readmissions. The ensemble approach effectively leveraged the strengths of different classifiers, significantly improving prediction accuracy.

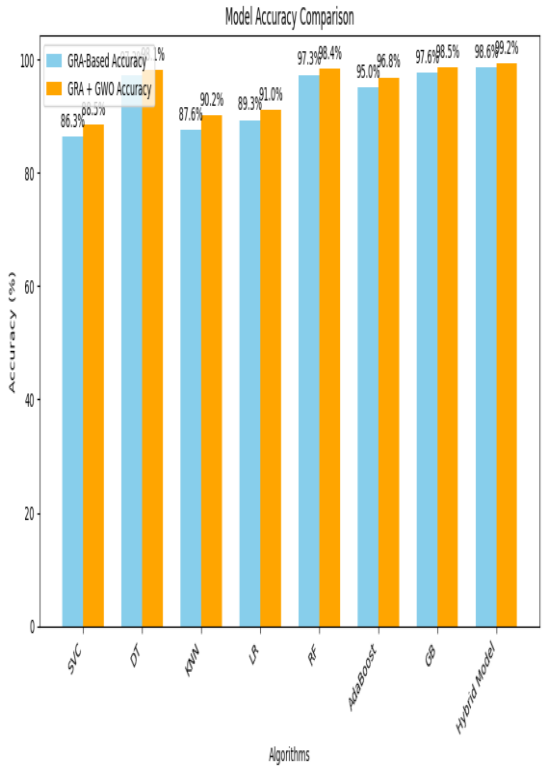
**METHODOLOGY**

In healthcare, the early prediction of diabetic patient readmission is crucial for improving patient outcomes and managing hospital resources efficiently. Given the chronic nature of diabetes and its associated complications, predicting which diabetic patients are at high risk of readmission is essential for early intervention and reducing hospital condition. In this methodology, we propose a hybrid framework that combines GRA for feature selection and Grey Wolf Optimization (GWO) for optimizing classification models. This framework helps predict the risk of readmission in diabetic patients and allows healthcare professionals to take preventive actions to avoid unnecessary condition. In the study, we proposed a methodology that leverages ML algorithms for predictive modelling of hospital condition among diabetic patients, with the ultimate goal of managing healthcare costs and improving patient outcomes. In addition to these traditional ML models, a deep learning approach was also considered by incorporating a LSTM model, to explore whether deep learning could outperform traditional ML models in terms of predictive performance. EHRs gathered from US hospitals provide the dataset used to train and test the models. It includes both continuous and categorical characteristics that are pertinent to diabetes patients and their condition to hospitals.

**RESULTS AND DISCUSSIONS**

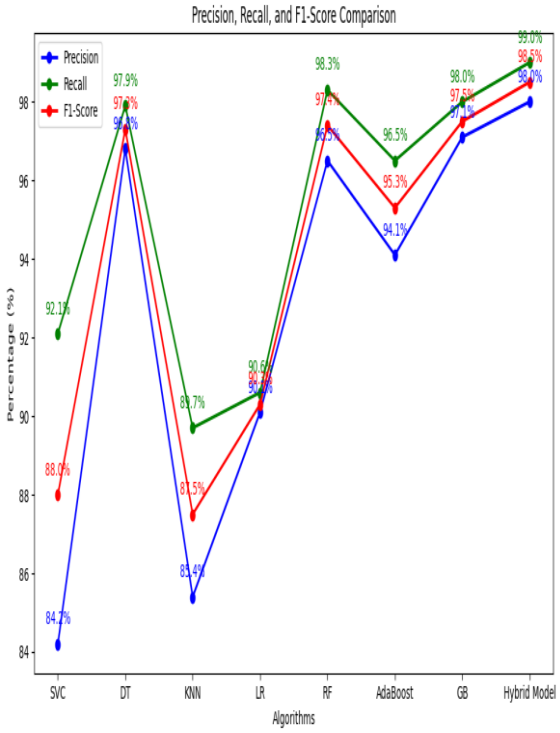
The GRA + GWO model consistently outperforms other classifiers, showing the effectiveness of using GRA for feature selection combined with GWO for optimization.

This study has developed a robust classification model for predicting diabetic patient condition, with a focus on classifying short-term and long-term condition. By leveraging GRA for feature selection and GWO for model optimization, the proposed methodology provides a high-accuracy solution for predicting condition.




**Graph 1: Accuracy Comparison between GRA and Traditional Classifiers**

Particularly when contrasted with conventional classifiers, the findings demonstrate significant gains in classification accuracy. The model's accuracy and efficiency are guaranteed by the combination of GRA and GWO, which makes it appropriate for real-world medical applications where precise forecasts may improve resource management and patient care.



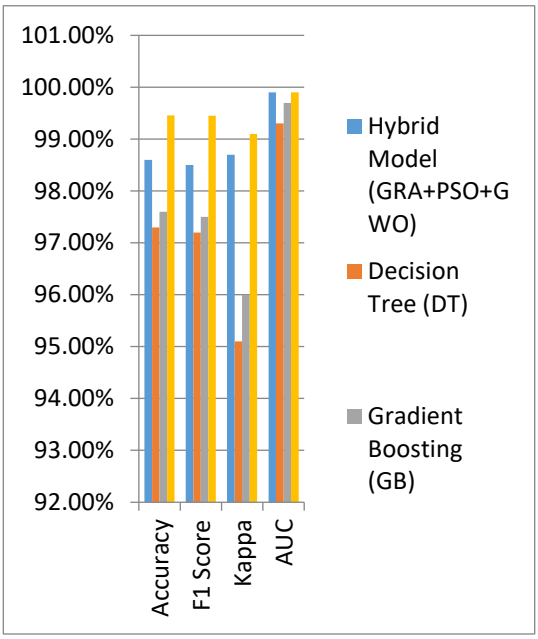

**Graph 2: Precision, Recall, and F1-Score Comparison Between GRA and Traditional Classifiers**

**Table 1: Accuracy Comparison of Classifiers**

| Method | Accuracy | F1 Score | Kappa | AUC |
|---|---|---|---|---|
| Hybrid Model (GRA+PSO +GWO) | 98.6 % | 98.5 % | 0.98 7 | 0.9 99 |
| Decision Tree (DT) | 97.3 % | 97.2 % | 0.95 1 | 0.9 93 |
| Gradient Boosting (GB) | 97.6 % | 97.5 % | 0.96 0 | 0.9 97 |
| SVM | 99.46 % | 99.4 5% | 0.99 1 | 0.9 99 |

The hybrid model demonstrates superior results by integrating GRA for feature selection and GWO for model optimization. The combination of these two advanced

techniques ensures high accuracy and robust predictions.




**Graph 3: performance of the ML classifiers**

This methodology, leveraging GRA for feature selection and GWO for optimization, forms a highly effective hybrid model for predicting the early readmission risk of diabetic patients. The model's ability to select relevant features and optimize hyper parameters results in superior predictive accuracy, making it a valuable tool in healthcare decision support systems.

## CONCLUSION

This research develops a smart decision support framework to forecast DM in diabetic patients. This is an area of prime concern in health management systems. Readmissions, especially for diabetic patients, come with high financial implications considering the sophisticated nature of chronic disease management and the strain on the healthcare system's capacity to minimize preventable admissions. Effectively managing these DM is critical from a patient care, resource management, and healthcare expenditure perspective. The research utilizes ML models, sophisticated feature selection, and optimization strategies to devise a powerful model for forecasting short and long-term readmissions. . This framework also enhances existing literature and research on predictive analytics of chronic illness along with highlighting the role of ML in healthcare through the power of data-driven decision support systems. To summarize, while this framework will help improve healthcare costs and the number of DM for high-risk diabetes patients, the outcomes will still be beneficial for lowering overall hospital admissions. It allows healthcare managers and providers to have better resource and patient care management using guided evidence.

## REFERENCES

1. *Aishwariya Dutta et.al (2022), "Early Prediction of Diabetes Using an Ensemble of Machine Learning Models", International Journal of Environmental Research and Public Health,ISSN:1660-4601,Volume.19, Issue.19*
2. *K Ramanan et.al (2024), "Predicting Diabetes Using Data Mining", International Journal of Creative Research Thoughts, ISSN: 2320-2882, Volume.12,Issue.12*
3. *Umair Butt et.al (2021), "Machine Learning Based Diabetes Classification and Prediction for Healthcare Applications", Journal of Healthcare Engineering, ISSN:2040-2309,vol.(7),Pages.1-17,DOI:10.1155/2021/9930985*
4. *Isfafuzzaman Tasin et.al (2023), "Diabetes prediction using machine learning and explainable AI techniques", Healthcare Technology Letters,ISSN:2053-3713,Volume.10,Issue.1-2*
5. *Y. Liu, B. Wang, L. Zhang, M. Li, and X. Chen (2020), "Gut microbiome fermentation determines the efficacy of*

*exercise for diabetes prevention," Cell Metabolism, vol. 31, no. 1, pp. 77-91.e5.*

6. *Wei, X., Jiang, F., Wei, F., Zhang, J., Liao, W., Cheng, S. (2017). An Ensemble Model for Diabetes Diagnosis in Large-scale and Imbalanced Dataset. International Journal of Medical Informatics, 107, 43–50. https://doi.org/ 10.1016/j.ijmedinf.2017.07.002*
7. *S. J. Healy, D. Black, et al. (2013), "Inpatient diabetes education is associated with less frequent hospital readmission among patients with poor glycemic control," Diabetes Care, vol. DC 130108, 2013.*
8. *Q. Li, R. Fu, J. Zhang, et al., " (2017) Label-Free Method Using a Weighted-Phase Algorithm To Quantitate Nanoscale Interactions between Molecules on DNA Microarrays," Anal. Chem., vol. 89, pp. 3501–3507, 2017*
9. *N.S. Gregory, J.J. Seley, S.K. Dargar, et al., "(2018) Strategies to Prevent Readmission in High-Risk Patients with Diabetes: The Importance of an Interdisciplinary Approach," Curr. Diab. Rep., vol. 18, no. 54, 2018.*
10. *N. Allaudeen, J. L. Schnipper, E. J. Orav, R. M. Wachter, and A. R. Vidyarthi, (2011) "Inability of providers to predict unplanned readmissions," Journal of General Internal Medicine, vol. 26, no. 7, pp. 771–776, 2011.*