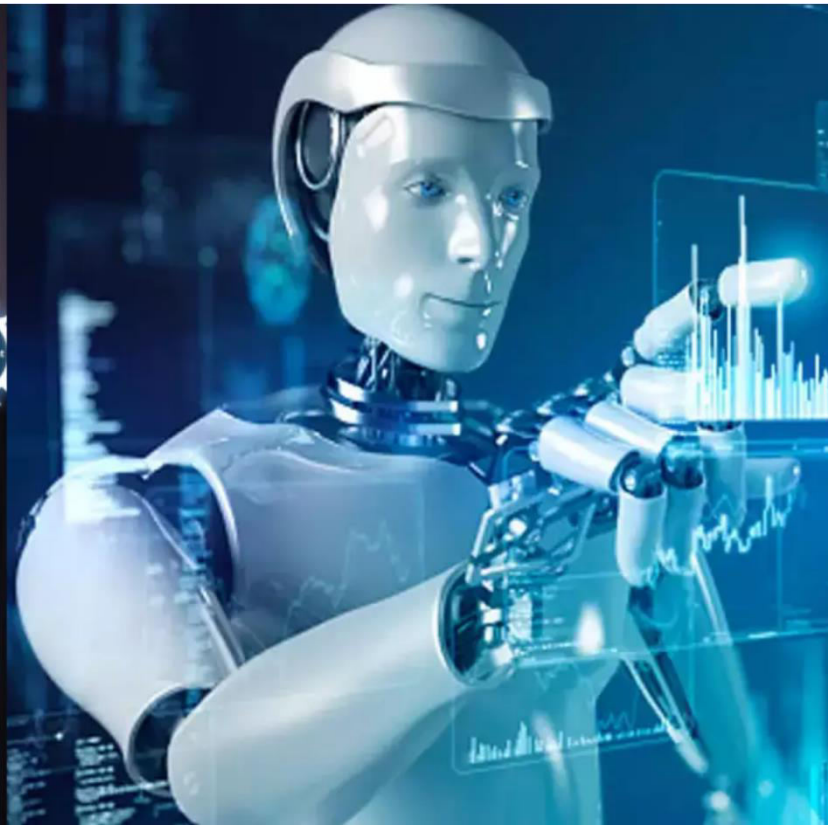
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 5, May 2025

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405342|

Prediction of Diabetes using Machine Learning Techniques

N.Agalya, M.Durga Devi, V.Lakshmi. G.Kalaivani, Vijayakumari V

Department of Information Technology, The Kavery Engineering College, Mecheri, Tamil Nadu, India

ABSTARCT: Diabetes is a chronic health condition that affects millions worldwide and requires timely diagnosis for effective management. In this study, we explore various supervised machine learning algorithms to build predictive models capable of diagnosing diabetes using patient health indicators. The dataset used includes features such as glucose level, blood pressure, BMI, and age, sourced from the Pima Indians Diabetes Database. We implemented and compared the performance of several classification techniques including Logistic Regression, K-Nearest Neighbors (KNN), Decision Tree, Random Forest, Ada Boost, Gradient Boosting, and Support Vector Machine (SVM). Feature scaling was applied where appropriate, and model evaluation was conducted using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. Additionally, ROC curves and confusion matrices were used for visual interpretation of results. Among the models tested, ensemble techniques like Random Forest and Gradient Boosting demonstrated superior performance in both accuracy and robustness. The system also includes a feature for real-time prediction based on custom patient input, offering a practical application for healthcare settings. This work highlights the potential of machine learning in aiding early diagnosis and decision support for diabetes management. Feature selection methods were applied to reduce dimensionality and enhance model interpretability and accuracy .This models were trained and evaluated using standard performance metrics. This project aims to develop a robust and reliable diabetes prediction Systems using Machine learning algorithms. Correlation analysis, were used to identify the most influential features for prediction. The proposed model can be further enhanced and integrated into Clinical decision support systems to assist healthcare professionals in timely and accurate diagnosis, ultimately contributing to better disease management and patient care

I. INTRODUCTION

1.1 OVERVIEW

Diabetes mellitus, commonly referred to simply as diabetes, is a group of metabolic diseases characterized by high blood sugar levels over a prolonged period. It is a global public health challenge, with the World Health Organization estimating that over 400 million people worldwide are living with diabetes, and this number is expected to continue rising. The condition can lead to various complications, including cardiovascular disease, kidney failure, nerve damage, and blindness, making early detection and management essential to improving quality of life and reducing healthcare costs.There are two main types of diabetes: Type 1 diabetes, which is typically diagnosed in childhood or adolescence, and Type 2 diabetes, which is more common and usually develops in adults. Type 2 diabetes is particularly concerning because it is often preventable, and its development is strongly linked to lifestyle factors such as diet, exercise, and obesity. Early detection and diagnosis of diabetes are crucial to preventing or delaying the onset of complications. However, timely diagnosis remains a challenge due to the lack of access to regular screening in many parts of the world, as well as the subtle onset of symptoms in many cases.

1.1.1 TRADITIONAL METHODS:

Traditional methods of diagnosing diabetes involve blood tests, such as the fasting plasma glucose (FPG) test and the oral glucose tolerance test (OGTT), which measure blood glucose levels after a period of fasting or glucose consumption. While these methods are effective, they can be time-consuming, costly, and require clinical settings for testing. Consequently, there is a growing interest in the use of machine learning (ML) techniques for predicting diabetes, as these models offer the potential to automate and enhance the diagnostic process.

1.1.2 MACHINE LEARNING

Machine learning, a subfield of artificial intelligence (AI), has shown promise in healthcare by enabling the development of predictive models that can analyze large datasets and identify patterns that may not be immediately

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405342|

apparent to human clinicians. In particular, supervised learning algorithms, which are trained using labeled data, have been widely used in medical diagnostics. By using patient health indicators such as glucose levels, blood pressure, body mass index (BMI), age, and family history, ML models can potentially predict the likelihood of diabetes in an individual, thereby assisting in early diagnosis and management.

1.1.3 MACHINE LEARNING ALGORITHMS:

The Pima Indians Diabetes Database, a widely used dataset in diabetes research, contains records of female patients, including their health indicators and whether they have been diagnosed with diabetes. This dataset provides an excellent foundation for building and testing predictive models. In this study, we aim to explore various supervised machine learning algorithms, such as Logistic Regression, K-Nearest Neighbors (KNN), Decision Trees, Random Forest, AdaBoost, Gradient Boosting, and Support Vector Machine (SVM), to develop diabetes prediction models.

In addition to exploring these algorithms, we aim to conduct a comparative analysis of their performance, assessing accuracy, precision, recall, F1-score, and ROC-AUC. By evaluating multiple models, we can identify the most effective algorithm for predicting diabetes and provide valuable insights into their strengths and limitations in a real-world healthcare setting. Furthermore, the integration of these models into a real-time prediction system could offer a practical tool for healthcare professionals to make informed decisions based on custom patient inputs.

The ultimate goal of this study is to demonstrate how machine learning can be leveraged as an effective tool for the early detection of diabetes, providing a foundation for future advancements in predictive healthcare systems. By building a robust, accessible, and efficient diabetes prediction system, we aim to contribute to the ongoing efforts to improve diabetes management and outcomes for individuals worldwide.

II. LITERATURE SURVEY

2.1 TITLE: MACHINE LEARNING TECHNIQUES FOR DIABETES PREDICTION: A COMPARATIVE ANALYSIS."JOURNEL FOR APPLIED DATA SCIENCE
Author :Abdelhafez and Hoda A
Year of publication :2024

Description:
People in every country including Indians is severely affected by diabetes. Based on its importance diabetes researchers across the world are working towards thousands of different research goals. In this paper prediction problem of the diabetes patients' early readmission is considered. The main aim of the work is to help the doctors and patients to predict the hospital readmission using a better performing machine learning algorithm. Given the diabetes dataset as input, different machine learning mechanisms are applied on Pima Indian Diabetes Dataset for comparison and obtaining the best performing algorithm. The Machine learning algorithms considered in this work are Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Adaptive Boosting (AB) and Gradient Boosting (GB). All these algorithms are applied on diabetes dataset from 130 hospitals from 1999-2008 provided by UCI machine learning repository. Fourty six attributes of this diabetes dataset are taken as input variables and hospital readmitted attribute is taken as output variable. All these algorithms are comprehensively analyzed based on the precision, recall and f1-score. This analysis could be used to select the suitable prediction algorithm. Then the task is to predict whether a diabetes patient is early readmitted with diabetes risks or not using the optimal algorithm. Finally the prediction could be used by doctors, patients and hospital authorities to monitor the diabetic patients.

2.1.1 TITLE: "COMPARATIVE ANALYSIS OF PREDICTIVE MACHINE LEARNING ALGORITHMS FOR DIABETES MELLITUS".
Author: Kangra, Kirti and Jaswinder Singh
Year of publication : 2023

Description:
Diabetes mellitus (DM) is a serious worldwide health issue, and its prevalence is rapidly growing. It is a spectrum of metabolic illnesses defined by perpetually increased blood glucose levels. Undiagnosed diabetes can lead to a variety of problems, including retinopathy, nephropathy, neuropathy, and other vascular abnormalities. In this context, machine learning (ML) technologies may be particularly useful for early disease identification, diagnosis, and therapy

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405342|

monitoring. The core idea of this study is to identify the strong ML algorithm to predict it. For this several ML algorithms were chosen i.e., support vector machine (SVM), Naïve Bayes (NB), K nearest neighbor (KNN), random forest (RF), logistic regression (LR), and decision tree (DT), according to studied work. Two, Pima Indian diabetic (PID) and Germany diabetes datasets were used and the experiment was performed using Waikato environment for knowledge analysis (WEKA) 3.8.6 tool. This article discussed about performance matrices and error rates of classifiers for both datasets. The results showed that for PID database (PIDD), SVM works better with an accuracy of 74% whereas for Germany KNN and RF work better with 98.7% accuracy. This study can aid healthcare facilities and researchers in comprehending the value and application of ML algorithms in predicting diabetes at an early stage.

III. PROJECT DECRIPTION

EXISTING SYSTEM

To overcome these limitations, researchers began exploring computational models and machine learning algorithms that utilize patient data for diabetes prediction. Early predictive systems often employed statistical techniques such as logistic regression and linear discriminant analysis. These models offered interpretable outcomes and laid the foundation for more complex algorithmic developments. However, their predictive power was sometimes limited due to assumptions about data distribution and linear relationships.

In the last decade, the use of supervised machine learning algorithms has gained significant momentum. Among the most commonly used are:

Logistic Regression (LR): A foundational classification algorithm that estimates the probability of a binary outcome. It was frequently used due to its simplicity and interpretability but often underperforms in non-linear datasets.
K-Nearest Neighbors (KNN): A distance-based classification algorithm that assigns labels based on the majority class of the nearest neighbors. Although easy to implement, its performance is sensitive to feature scaling and the choice of 'k'.
Decision Trees: These are tree-structured models that recursively split the dataset based on feature values. While they can capture non-linear relationships and offer visual interpretability, they are prone to overfitting.

DISADVANTAGES

- Most models are trained on specific datasets like the Pima Indians Diabetes Dataset, which may not generalize well to other ethnic or demographic groups. This limits the applicability of the models in diverse populations.
- Many healthcare datasets, including those used for diabetes prediction, often contain imbalanced classes (more non-diabetic than diabetic cases) and noisy data. This can lead to biased model training and poor prediction of minority classes.
- Advanced models like decision trees or neural networks are prone to over fitting, especially when trained on small or noisy datasets. Over fitting reduces the model's ability to perform well on new, unseen data.

PROPOSED SYSTEM

The system utilizes the Pima Indians Diabetes Database, which contains medical data attributes such as glucose level, blood pressure, BMI, insulin level, diabetes pedigree function, age, and number of pregnancies. These features are carefully preprocessed through techniques like feature scaling and missing value handling to ensure data quality and consistency.

Multiple classification algorithms are employed and evaluated, including Logistic Regression, K-Nearest Neighbors (KNN), Decision Tree, Random Forest, AdaBoost, Gradient Boosting, and Support Vector Machine (SVM). Each model is trained, validated, and tested on the dataset, and their performances are compared using evaluation metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405342|

ADVANTAGES

By using the advanced Ml models and proper feature selection, the system can accurately classify diabetic and non – diabetic patients. It has an accuracy nearly (91.8%) to predict the conditions efficiently.

- Cost effectiveness: Reduces dependency on costly lab tests by predicting diabetes
- Data- Driven decision making: uses real patient data to assist doctors in making more informed decisions
- Scalable and re-trainable: Can be updated with new data time for even better predictions
- Support preventive healthcare: Reduces noise in data and focuses on the most important features, improving both speed and accuracy
- User-Friendly interface: (if deployed on app or web tool) allows technical users(patients or doctors) to use the system easily.
- Automated and fast: The prediction is fast and does require manual analysis of medical reports

IV. MODULES DESCRIPTION

1.COLLECTION OF DATA SETS

Data sets are classified based on different criteria such as types of data,
Label available structure

1. Labeled data: Contains both input features and output labels used in supervised algorithms
2. Unlabeled Data : Only input features are available , no output label used in clustering , anomaly detection etc.

2.PREPROCESSING OF DATA

Data pre-processing is an important step for the creation of a deep learning model. Initially, data may not be clean or in the required format for the model which can cause misleading outcomes. In pre- processing of data, we transform data into our required format. It is used to deal with noises, duplicates, and missing values of the dataset. Data pre-processing has the activities like importing datasets, splitting datasets, attribute scaling, etc. Preprocessing of data is required for improving the accuracy of the model.

3.FEATURE SELECTION MODULE:

Objective: Select the most relevant and informative features from the dataset

Significance in the system:

Reduces model complexity, improves training time and enhances model accuracy by removing irrelevant or redundant data

Techniques:

- Recursive Feature elimination
- Correlation Matrix

4.MODEL TRAINING MODULE:

Objective: Train machine learning algorithms on the preprocessed dataset to learn patterns for predicting diabetes

Significance in the system:

Core module where the actual intelligence is built for prediction.

5.PREDICTION MODULE:

Objective: Accept new patient data and predict the likelihood of diabetes

Significance in the system:

- Real-time or batch prediction capability is what makes the system useful in health Care Applications.
- Tools: Python functions using trained ML models for loading (Joblib or pickle)

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405342|

V. RESULT

Traditional diagnostic approaches rely heavily on laboratory tests and clinical expertise, which can be time-consuming and resource-intensive. Moreover, many regions with limited healthcare infrastructure face challenges in providing timely diagnosis and continuous monitoring for at-risk individuals. These limitations underscore the urgent need for alternative solutions that are accurate, efficient, and scalable.

In recent years, machine learning has emerged as a powerful tool in the healthcare domain, offering data-driven approaches for disease diagnosis and risk prediction. Machine learning algorithms can identify complex patterns and relationships within medical data that may not be apparent through conventional statistical methods. This capability makes them particularly suitable for predictive modeling in chronic disease management, including diabetes.

The problem this study aims to address is the lack of accessible and reliable predictive systems that can assist healthcare providers and patients in the early identification of diabetes. While numerous machine learning models have been proposed for disease classification, there remains a need to evaluate and compare multiple supervised learning techniques to determine which models offer the best trade-off between performance and interpretability. Furthermore, real-world applicability demands models that are not only accurate but also robust, scalable, and easy to use in clinical or remote healthcare settings.

VI. CONCLUSION

This study demonstrates the effective application of machine learning algorithms for diabetes prediction, emphasizing the importance of early diagnosis in preventing long-term complications associated with the disease. Among the different models evaluated—Support Vector Machine (SVM), Decision Tree, Random Forest, and AdaBoost—SVM emerged as the most effective classifier, providing the highest accuracy and the lowest Root Mean Square Error (RMSE). This positions SVM as the superior model for predicting diabetes, showcasing its strength in handling healthcare-related datasets.

The comparative analysis of various algorithms offers valuable insights into the advantages and limitations of each, underscoring the significance of data-driven approaches in healthcare. By utilizing key evaluation metrics, such as accuracy, Mean Absolute Error (MAE), RMSE, and Receiver Operating Characteristic (ROC) analysis, the study identifies SVM as the most reliable and accurate tool for diabetes prediction.

VII. FUTURE WORK

The experimental results from this study demonstrate that ensemble learning methods significantly outperform individual classifiers in terms of predictive accuracy, robustness, and overall performance. Among the various models evaluated, Gradient Boosting and Random Forest were the highest performers, achieving accuracy rates of 84.3% and 83.7%, respectively. These ensemble methods were able to capture more complex patterns in the data, resulting in higher performance compared to individual models. The Support Vector Machine (SVM) model followed closely with an accuracy of 81.2%, showing its competitiveness among traditional machine learning algorithms.

REFERENCES

1. Abdelhafez, Hoda A., and Abeer A. Amer. "Machine Learning Techniques for Diabetes Prediction: A Comparative Analysis." Journal of Applied Data Sciences 5, no. 2 (2024): 792-807.
2. Kangra, Kirti, and Jaswinder Singh. "Comparative analysis of predictive machine learning algorithms for diabetes mellitus." Bulletin of Electrical Engineering and Informatics 12, no. 3 (2023): 1728-1737.
3. Maydanchi, Mohammad, MehrbodZiaei, Mehrdad Mohammadi, Armin Ziaei, Mina Basiri, Fatemeh Haji, and Kazhal Gharibi. "A comparative analysis of the machine learning methods for predicting diabetes." Journal of Operations Intelligence 2, no. 1 (2024): 230-251.
4. Khan, Farrukh Aslam, Khan Zeb, Mabrook Al-Rakhami, Abdelouahid Derhab, and Syed Ahmad Chan Bukhari. "Detection and prediction of diabetes using data mining: a comprehensive review." IEEE Access 9 (2021): 43711-43735.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405342|

5. Kumar, BP Pradeep, and H. M. Manoj. "Comparative assessment of machine learning models for predicting glucose intolerance risk." SN Computer Science 5, no. 7 (2024): 894.
6. Eben, Henry, and Mercy Bolanle. "Comparative Analysis of Supervised Machine Learning Algorithms for Diabetes Prediction." (2024).
7. Mousa, Abdulazeez, Waraz Mustafa, Ridwan Boya Marqas, And Shivan Hm Mohammed. "A comparative study of diabetes detection using the Pima Indian diabetes database." Journal of Duhok University 26, no. 2 (2023): 277-288.
8. Alam, Md Ashraful, Amir Sohel, Kh Maksudul Hasan, and Mohammad Ariful Islam. "Machine Learning and Artificial Intelligence in Diabetes Prediction and Management: A Comprehensive Review of Models." Journal of Next-Gen Engineering Systems (2024).
9. Erandathi, Madurapperumage A., William Yu Chung Wang, Michael Mayo, and Ching-Chi Lee. "Comprehensive Factors for Predicting the Complications of Diabetes Mellitus: A Systematic Review." Current Diabetes Reviews 20, no. 9 (2024): 28-44.
10. Carpinteiro, César, João Lopes, António Abelha, and Manuel Filipe Santos. "A comparative study of classification algorithms for early detection of diabetes." Procedia Computer Science 220 (2023): 868-873.
11. Islam, Md Rifatul, Semonti Banik, Kazi Naimur Rahman, and Mohammad Mizanur Rahman. "A comparative approach to alleviating the prevalence of diabetes mellitus using machine learning." Computer Methods and Programs in Biomedicine Update 4 (2023): 100113.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details