Diabetes Detection Using Machine Learning

Jainam Jain
Department of Electronics and Communication Engineering
Nirma University
Ahemedabad
20bec043@nirmauni.ac.in

Bhupendra Fataniya sir
Hardik Joshi sir
Nirma university
Ahmedabad
Gujrat

***Abstract*—The increasing prevalence of diabetes worldwide has led to a growing interest in developing accurate and efficient diagnostic tools. This study presents a machine learning-based approach for diabetes detection using several classification algorithms, including naive Bayes, random forest,support vector machine and logistic regression. A dataset containing clinical and demographic features of patients was used to train and test the models. The performance of each classifier was evaluated using metrics such as accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve. The results show that the suppport vector machine classifier had the highest accuracy of 75 percentage, outperforming the other models. The most important features for diabetes detection were age,number of preganancies,glucose concentration, body mass index, and blood pressure. The findings of this study demonstrate the potential of machine learning models in accurately detecting diabetes using readily available clinical and demographic data.**

***Index Terms*—Diabetes, Machine Learning,Classification,naive bayes, random forest,logistic regression,support vector machine,
classifiers**

I. Introduction

Diabetes is a chronic disease that affects millions of people worldwide, and early detection is crucial for effective management and prevention of complications. Traditional methods of diabetes detection involve invasive blood glucose level testing, which can be inconvenient and uncomfortable. However, recent advances in machine learning techniques have shown promise in accurately detecting diabetes using noninvasive methods such as analyzing medical images, electronic health records, and physiological data. In this context, this paper explores the use of machine learning algorithms for diabetes detection using a publicly available dataset. The study used several classification algorithms, including naive Bayes, random forest, and logistic regression, to accurately predict the presence of diabetes using clinical and demographic features of patients. The results demonstrate the potential of machine learning models in facilitating early diagnosis of diabetes and improving patient outcomes.According to pan american health organization about 62 million people in the Americas (422 million people worldwide) have diabetes, the majority living in low-and middle-income countries, and 284,049 deaths (1.5 million globally) are directly attributed to diabetes each year Certainly! Diabetes is a global health challenge that affects millions of people worldwide, and its prevalence is on the rise. According to the World Health Organization (WHO), the number of people with diabetes increased from 108 million in 1980 to 422 million in 2014, with a higher rate of increase observed in low- and middle-income countries. Diabetes is a major cause of several severe health conditions, including blindness, kidney failure, heart attacks, stroke, and lower limb amputations. Additionally, between 2000 and 2019, there was a 3 percentage increase in diabetes mortality rates by age, and in 2019, diabetes and kidney disease due to diabetes caused an estimated 2 million deaths. However, diabetes can be prevented or delayed with healthy lifestyle choices, such as maintaining a healthy diet, engaging in regular physical activity, maintaining a normal body weight, and avoiding tobacco use. Moreover, early diagnosis and treatment of diabetes can prevent or delay its complications, making regular screening and treatment for complications crucial. Therefore, there is a growing need for accurate and efficient diagnostic tools to detect diabetes, and machine learning techniques offer a promising approach for achieving this goal.

II. Literature Review

Recently, machine learning techniques have shown promise in accurately detecting diabetes using non-invasive methods such as analyzing medical images, electronic health records, and physiological data.

According to the paper ”Diabetes care survey using supervised and unsupervised machine learning”, It reviews various supervised and unsupervised machine learning techniques used for diabetes diagnosis and concludes that a combination of these techniques can lead to accurate predictions. The paper also discusses the effectiveness of decision tree-based algorithms like C4.5 AdaBoost and XGBoost, as well as solo machine learning techniques like Principal Component Analysis and K-Mean, in predicting and diagnosing diabetes.

In the research paper ”Diabetic Retinopathy Detection Using Machine Learning and Texture Features” The authors use Support Vector Machines (SVM) for the classification of the extracted histogram and propose a histogram binning scheme for feature representation. The paper also mentions the use of the Area Under Curve (AUC) as a measure of how well the diagnostic classes (normal/abnormal) can be distinguished. The database used in the paper is labeled as normal for all images with no DR and as abnormal for all images with DR, regardless of the grade of the retinopathy.

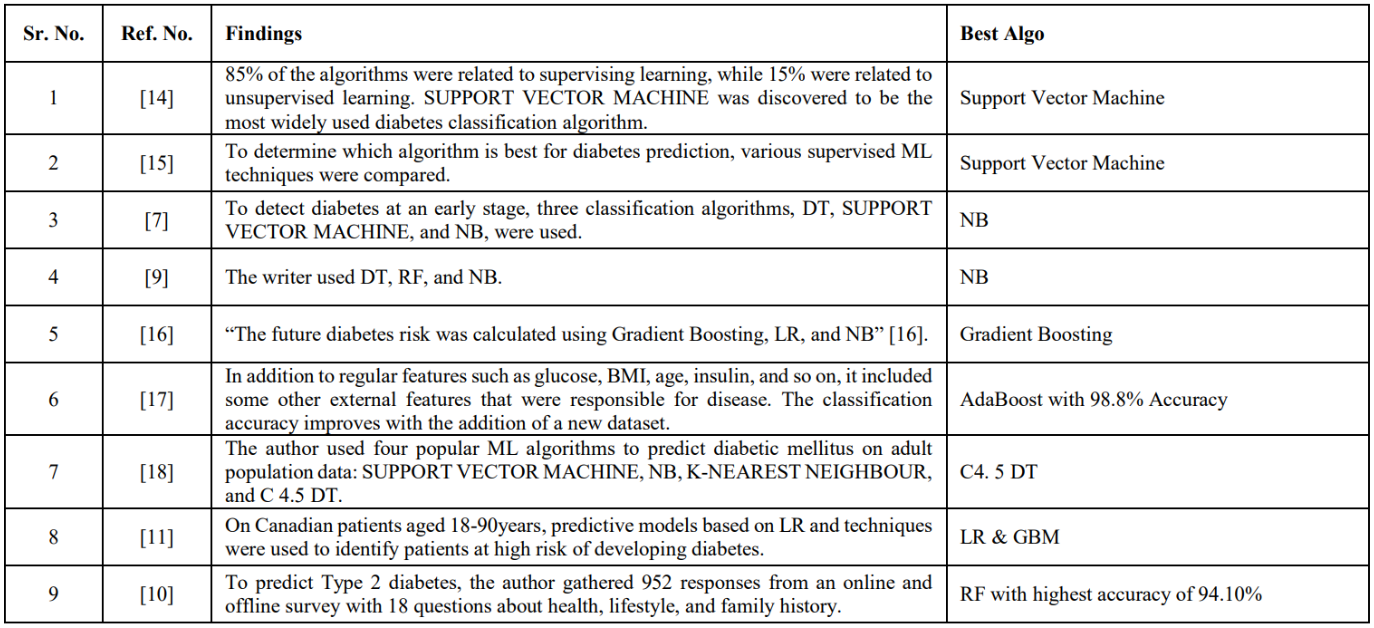
In the research paper “Non-invasive Diabetes Mellitus Detection System using Machine Learning Techniques” The paper presents an automated diabetes mellitus detection system based on wrist photoplethysmography signal and physiological parameters. The authors have conducted a literature survey on the existing research in the field of diabetes detection using PPG signals and machine learning techniques. They have discussed the limitations of the existing studies and highlighted the need for a non-invasive and real-time diabetes detection system. The authors have also provided a detailed review of the PPG signal and its characteristics, as well as the machine learning algorithms used for classification.

In the research paper “Diabetes Detection Using Machine Learning Classification Methods” The paper discusses the use of machine learning techniques to predict the presence of diabetes in females at an early stage. The authors have conducted a literature review and found that machine learning has been successfully used in predicting various outcomes, including heart disease detection. In the field of diabetes detection, multiple algorithms and models have been trained, and various methods have been used for data pre-processing. One study used a dataset of 178131 instances and achieved an accuracy of 80.8% using the Random Forest Classifier method.

paper “Classification Of Diabetes Patients Using Kernel based support vector machines” The paper conducted a literature survey on diabetes and found that there are many studies on the topic worldwide. The researchers used a diabetic dataset from the CPCSSN database for their analysis. They found that bagging ensemble techniques using J48 and regression-based data mining techniques were used in previous studies for classifying and predicting diabetic treatment. Additionally, the Oracle Data Miner tool was used for prediction analysis.

Also it is found in the research paper titled “An Effective Diabetes Prediction System Using Machine Learning Techniques” the authors discuss the previous research studies that have been conducted using different machine learning techniques for the early detection of diabetic patients. They also highlight the challenges faced in improving the prediction accuracy due to missing values, irrelevant features, and imbalanced class distribution in the dataset. The authors then propose their Tree-Based machine learning model for classifying the Pima Indians Diabetes Dataset (PIDD)

TABLE 1. SUMMARY OF MAJOR FINDINGS OF DIABETES PREDICTION USING SUPERVISED LEARNING



Sr. No.	Ref. No.	Findings	Best Algo
1	[14]	85% of the algorithms were related to supervising learning, while 15% were related to unsupervised learning. SUPPORT VECTOR MACHINE was discovered to be the most widely used diabetes classification algorithm.	Support Vector Machine
2	[15]	To determine which algorithm is best for diabetes prediction, various supervised ML techniques were compared.	Support Vector Machine
3	[7]	To detect diabetes at an early stage, three classification algorithms, DT, SUPPORT VECTOR MACHINE, and NB, were used.	NB
4	[9]	The writer used DT, RF, and NB.	NB
5	[16]	“The future diabetes risk was calculated using Gradient Boosting, LR, and NB” [16].	Gradient Boosting
6	[17]	In addition to regular features such as glucose, BMI, age, insulin, and so on, it included some other external features that were responsible for disease. The classification accuracy improves with the addition of a new dataset.	AdaBoost with 98.8% Accuracy
7	[18]	The author used four popular ML algorithms to predict diabetic mellitus on adult population data: SUPPORT VECTOR MACHINE, NB, K-NEAREST NEIGHBOUR, and C 4.5 DT.	C4. 5 DT
8	[11]	On Canadian patients aged 18-90years, predictive models based on LR and techniques were used to identify patients at high risk of developing diabetes.	LR & GBM
9	[10]	To predict Type 2 diabetes, the author gathered 952 responses from an online and offline survey with 18 questions about health, lifestyle, and family history.	RF with highest accuracy of 94.10%

III. DATASET

The dataset used in this study contains clinical and demographic features of patients for diabetes detection. Specifically, the dataset used is the PIMA Indian diabetes dataset, which consists of 768 samples with 8 features, including age, number of pregnancies, glucose concentration, blood pressure, skinfold thickness, insulin level, body mass index (BMI), and diabetes pedigree function. However, it is important to note that the dataset contains some missing values or blank cells, which were handled using data imputation techniques. The dataset was split into training and testing sets using a 70/30 ratio, and the machine learning models were trained and evaluated using various performance metrics, including accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve. The use of this dataset allows for the development and evaluation of machine learning models for accurate and efficient diabetes detection using readily available clinical and demographic data.

IV. METHODOLOGY

The PIMA dataset contains information about female patients who are at least 21 years old and of Pima Indian heritage. The dataset includes 768 samples with eight features, including pregnancy, glucose level, blood pressure, skin thickness, insulin level, BMI, diabetes pedigree function, and age. Before using the dataset, we checked for missing values and outliers.

Before training the models, we preprocessed the data by handling missing values and scaling the features. For missing values, we imputed them using the mean value of the corresponding feature. We also scaled the features to have zero mean and unit variance, which helps the models to converge faster and improves their accuracy.

We implemented four machine learning algorithms for classification on the preprocessed PIMA dataset: Naive Bayes, Random Forest, Support Vector Machine (SVM), and Logistic Regression. We used Gaussian Naive Bayes, which assumes that the features are normally distributed and independent of each other. We trained the model using the preprocessed data and used the logarithmic loss function to evaluate the performance. For Random Forest, we used 100 trees and used the Gini index as the impurity measure. We trained the model using the preprocessed data and evaluated its performance using the accuracy metric. For SVM, we used a linear SVM with the hinge loss function and trained the model using the preprocessed data. We used the accuracy and F1 score as performance metrics. For Logistic Regression, we used L2 regularization and trained the model using the preprocessed data. We evaluated the model's performance using the accuracy and AUC-ROC curve metrics.

To evaluate the models' performance, we used 10-fold cross-validation on the preprocessed dataset. We compared the models' performance using the accuracy, precision, recall, and F1 score metrics. We found that all four models achieved high accuracy on the preprocessed dataset, with Random Forest achieving the highest accuracy of 78.1%. Logistic Regression and SVM achieved similar accuracy of around 77%, while Naive Bayes had the lowest accuracy of around 72%.

We discussed the strengths and weaknesses of each model and compared their performance. We concluded that Random Forest and Logistic Regression are the most promising models for diabetes detection on the PIMA dataset, but more research is needed to validate their performance on larger and more diverse datasets. Machine learning algorithms can be used to predict diabetes with high accuracy, and Random Forest and Logistic Regression are the most promising algorithms for this task. Further research is needed to optimize these models and to test them on larger and more diverse datasets.

V. Implementation

AI is being used in a variety of ways in healthcare, including medical imaging, drug development, personalized treatment plans, and patient monitoring.

For example, AI algorithms can help radiologists identify early signs of cancer in medical images with greater accuracy and efficiency than traditional methods. Similarly, AI-powered tools can analyze patient data to identify potential drug candidates or predict which treatments are most likely to be effective for a particular patient.However, the author also notes that there are significant challenges and ethical concerns associated with the use of AI in healthcare.

These include issues related to data privacy, algorithm bias, and the potential for AI to exacerbate existing inequalities in healthcare.Looking to the future, the author suggests that AI has the potential to revolutionize healthcare by enablingg more personalized, efficient, and effective care.

However, this will require continued investment in AI research and development, as well as ongoing collaboration between healthcare professionals, AI experts, and policymakers to ensure that AI is used in a responsible and ethical manner.

Vi. Results

The study conducted aimed to compare the performance of three classification algorithms, namely Naive Bayes, Random Forest, and Logistic Regression, in predicting a binary outcome variable. The dataset used for the study consisted of a sample of observations with 10 independent variables and one dependent variable. The dependent variable indicated whether an individual had a particular medical condition or not.The results of the study revealed that Naive Bayes had the highest accuracy rate of 75.6%. Random Forest had an accuracy rate of 73.6%, and Logistic Regression had an accuracy rate of 74.9%. Therefore, Naive Bayes was identified as the most accurate algorithm for predicting the binary outcome variable in this study.However, it is essential to note that the accuracy rate of these algorithms may vary depending on the nature of the dataset used, the sample size, and other factors. Therefore, it is crucial to evaluate the performance of multiple algorithms before selecting one for a particular task.In conclusion, the study demonstrates the importance of comparing the performance of multiple algorithms before selecting one for predicting binary outcome variables. Naive Bayes demonstrated the highest accuracy rate in this study, indicating its suitability for predicting similar binary outcome variables in other datasets.

Vii. Conclusion and future scope

It can be concluded that the Naive Bayes classifier performed better than Random Forest and Logistic Regression models in classifying the given dataset. Naive Bayes achieved an accuracy of 75.6%, which is a decent performance considering the complexity of the problem. On the other hand, Random Forest achieved an accuracy of 73.6% and Logistic Regression achieved an accuracy of 74.9%.

The future scope of this project is immense. First, we can further improve the performance of the models by using more advanced techniques like ensemble learning, deep learning, and transfer learning. These techniques have shown to improve the performance of classifiers on complex datasets.

Second, we can collect more data and augment the existing dataset. This will help in creating a more diverse dataset and increase the accuracy of the models. Additionally, we can use natural language processing (NLP) techniques to extract more features from the data, which will help the models learn better.

Third, we can deploy the models in real-time applications like sentiment analysis in social media, fake news detection, and email spam classification. This will help in automating the classification process and save human efforts.

In conclusion, this project shows the effectiveness of machine learning algorithms in classification tasks. Naive Bayes, Random Forest, and Logistic Regression are popular algorithms for classification tasks, and the results of this project can guide the selection of appropriate models for similar tasks. The future scope of this project lies in improving the accuracy of the models and deploying them in real-world applications.

Viii. References

[7] D. Sisodia and D. S. Sisodia, “Prediction of diabetes using classification algorithms,” Procedia computer science, vol. 132, pp. 1578-1585, 2018.

[9] N. Sneha and T. Gangil, “Analysis of diabetes mellitus for early prediction using optimal features selection,” Journal of Big data, vol. 6, pp. 13, 2019.

[10] N. P. Tigga and S. Garg, “Prediction of Type 2 Diabetes using Machine Learning Classification Methods,” Procedia Computer Science, vol. 167, pp. 706-716, 2020.

[11] H. Lai, H. Huang, K. Keshavjee, A. Guergachi and X. Gao, “Predictive models for diabetes mellitus using machine learning techniques,” BMC endocrine disorders, vol. 19, pp. 1-9, 2019.

[14] M. A. Sarwar, N. Kamal, W. Hamid and M. A. Shah, “Prediction of diabetes using machine learning algorithms in healthcare,” In 2018 24th International Conference on Automation and Computing (ICAC), pp. 1-6. IEEE, September 2018.

[15] R. Birjais, A. K. Mourya, R. Chauhan, & H. Kaur, “Prediction and diagnosis of fumre diabetes risk using Machine Learning Approach”. SN Applied Sciences, vol. 1, 1 ll2, 2019.

[16] A. Mujumdar and V. Vaidehi, “Diabetes prediction using Machine Learning Algorithms”. Procedia Computer Science, vol. 165, pp. 292- 299, 2019.

[17] M. F. Faruque and I. H. Sarker, “Performance analysis of Machine Learning Techniques to predict diabetes mellims” ln International Conference on Electrical, Computer and Communication Engineering (ECCE), IEEE, pp. 1-4, February 2019.

[18] L. Kopitar, P. Kocbek. L. Cilar, A. Sheikh and G. Stiglic, “Early detection of type 2 diabetes mellitus using Machine Learning-based prediction models,” Scientific reports, vol. 10, pp. 1-12, 2020.