ISSN 2063-5346 *Section A-Research paper*

Early Detection and Segmentation of Diabetes Based on Machine Learning

[1] Joy Chandra Talukder, [2] Sooraj Sreekumar, [3] Nandlal Das, [4] Pandilya Trivedi, [5] MD Najmus Sakib, [6] Premananda Sahu.

Email: [1] joychandratalukder@gmail.com, [2] s007rajs@gmail.com, [3] nandarya61@gmail.com, [4] arocks061@gmail.com, [5] cadetsakib152@gmail.com , [6] premananda.29813@lpu.co.in

[1, 2, 3, 4, 5, 6] School of Computer Science and Engineering, Lovely Professional University, Phagwara, Punjab, India.

Abstract- Worldwide, millions of people struggle with the chronic illness of diabetes. Early detection and treatment of diabetes can prevent or delay complications. Both diabetes diagnosis and prediction have greatly benefited from the use of ML techniques. With the use of a dataset of clinical and demographic characteristics, we present in this work a ML-based method for diabetes prediction. Comparisons are made between Decision Tree, RF, Naive Bayes, Boosting Algorithm, and SVM in terms of their efficacy. Our findings show that the Naive Algorithm outperformed the other algorithms, with an accuracy of 76%. The most significant indicators of diabetes are, according to our research, pregnancies, BP, skin thickness, insulin, a family history of diabetes, age, glucose levels and Body Mass Index (BMI). Our method may be used as a screening tool for early diabetes identification, enabling prompt management and intervention.

Keyword- Diabetes, Random Forest, Machine Learning, Naive Bayes, SVM.

I. Introduction

Diabetes, a chronic metabolic disorder that, if left untreated, can lead to serious health complications, is characterized by high blood glucose levels. To successfully control the condition and avoid long-term consequences, early identification and classification of diabetes are crucial. Machine learning techniques have recently demonstrated considerable potential in the detection and segmentation of diabetes from medical imaging data, including retinal pictures and magnetic resonance imaging (MRI) scans. These methods are capable of analyzing vast volumes of data and locating tiny patterns that the human eye could miss.

Many individuals worldwide are afflicted by diabetes, a chronic metabolic illness. It is characterized by persistently high blood sugar levels, which result from the body's inadequate insulin production or ineffective insulin usage. The increased frequency of diabetes over the past few decades has made it a significant public health problem.

Early detection and segmentation of diabetes are crucial for effective management of the disease and preventing complications. However, traditional diagnostic methods such as blood glucose tests, HbA1c tests, and oral glucose tolerance tests are invasive, time-consuming, and expensive. Nowadays, machine learning promises us about early detection and segmentation of diabetes.

One of two categories—supervised or unsupervised—describes the majority of machine learning algorithms. Unsupervised algorithms can infer results from datasets, whereas supervised algorithms use historical data to predict outcomes for new or unobserved data [1]. Large datasets can be analyzed by machine learning algorithms, and these algorithms can spot patterns that human experts might miss. These algorithms can be trained using patient data to identify and categorize diabetes depending on a number of variables like age, weight, family history, and other medical conditions. Additionally, ML algorithms can also segment diabetic lesions from medical images, providing physicians with critical information for diagnosis and treatment.

One of the most significant advantages of ML-based approaches for diabetes detection and segmentation is their ability to provide personalized care. ML algorithms can analyze individual patient data and provide tailored treatment recommendations, improving patient outcomes and reducing healthcare costs. Additionally, these methods can be used to

forecast a person's chance of acquiring diabetes in the future, enabling preventive measures to be taken.

Despite the potential benefits of ML-based approaches for diabetes detection and segmentation, there are various issues that need to be resolved. The requirement for vast and varied datasets to accurately train the ML algorithms is one of the major problems. The proper expectation model would need extra pertinent information to make it progressively precise [2]. Additionally, there is a need for standardized protocols for data collection, annotation, and evaluation.

In this project, we target to explore and develop ML techniques for early detection and segmentation of diabetes, with the goal of improving patient outcomes and quality of life. We will review the state-of-the-art ML-based approaches for early detection and segmentation of diabetes. We will discuss the various datasets, features, and algorithms used in these approaches and highlight their strengths and limitations. We will also explore future directions and challenges in this field, emphasizing the need for collaboration between researchers, clinicians, and policymakers to improve the quality of diabetes care.

II. Literature Survey

Alehegn M. et al. [3] In order to prevent human deaths, the article covers the use of Data Mining Techniques (DMTs) to forecast medical datasets early on. The primary killer in the world today, Diabetes Disease (DD), is the subject of the study. To recognize diabetes this proposed system uses 768 records from the Pima Indian Diabetes Data Set along with a proposed Ensemble Method (PEM), Naive Net, SVM and Decision Stump. With a high accuracy of 90.36 percent, the PEM model integrates many methodologies and procedures into one. With a high accuracy of 90.36 percent, the PEM model integrates many methodologies and procedures into one. Sisodia D. and Sisodia DS. [4] This research paper describes that they used three ML algorithms to forecast the beginning of the stage of diabetes, which are SVM, Decision Tree and naive Bayes. The Pima Indians Diabetes Database (PIDD) was used, which was available from the UCI ML repository, and was utilized in the study to compare how well various algorithms performed on metrics including recall, precision, and accuracy. The study described in the paper employs Decision Tree, SVM, and Naive Bayes, three ML classification algorithms, to recognize diabetes at the beginning of the period. According on the findings, the Naive Bayes classification method has a 76.30% accuracy rate.

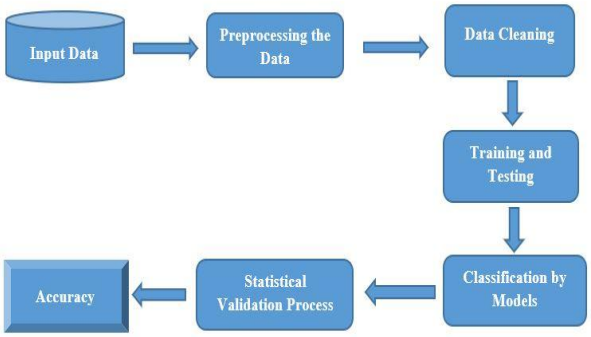
Mujumdar A. and Vaidehi V. [5] In addition to the usual variables like glucose, body mass index, age, and insulin, this study suggests a diabetes prediction model that takes into account extraneous variables. On the dataset, the study applies various machine learning techniques and discovers that logistic regression has the greatest accuracy (96%). The AdaBoost classifier is found to be the top model, with an accuracy of 98.8%, in the study, which also employs a pipeline model to further increase the accuracy. The study also implies that the model might be expanded to forecast the probability that those without diabetes will get the disease in the future. Krishnamoorthi R et al. [6] The use of ML methods to forecast diabetes is covered in the article. It emphasizes the value of early diagnosis and the advantages it provides for patients. The study presents various existing ML classification models for predicting diabetes based on their accuracy, and one such model is trained on the PIDD dataset. In addition, association rule mining was used to discover that glucose and BMI strongly reciprocity with diabetes and that Logistic Regression (LR) performed better than other machine learning algorithms. The LR model achieved a ROC value of 86%. The study's limitation was the use of a structured dataset; future research will take unstructured data into account.

Zou Q et al. [7] Hospital physical examination data from Luzhou, China, which consisted of 14 variables, were used in the study described in the article to forecast diabetes mellitus using decision trees, RF, and neural networks. In the study, the models were assessed using fivefold cross-validation. The researchers used PCA and mRMR to analyze data from 68994 healthy people and diabetic patients in order to reduce dimensionality. Islam S.M.S et al. [8] The identification of the microaneurysms (MAs) in this study allowed for the development of a special convolutional neural network (CNN) with strong educational experiences for the diagnosis at primary level of Diabetic Retinopathy (DR), which are the first signs of DR, and categorizing images of the retinal fundus into five groups. The network demonstrated state-of-the-art performance on severity

grading with a quadratic weighted an AUC value of 0.844 and a kappa value of 0.851 on the largest publicly available Kaggle dataset for diabetic retinopathy. Additionally, the network demonstrated its effectiveness by achieving a susceptibility of 98% and a specificity above 94% for the diagnosis at primary level. In terms of processing time and space, the suggested architecture is straightforward and effective.

III. Methodology

This aspect of the work primarily focuses on the diabetes prediction shown in Figure 1.



The graphic above has been divided into the subsequent processes.

- The data from Kaggle was initially used.
- The data will be needed for the preparation process.
- After preprocessing data need to be clean
- After cleaning data set need to be train and test for use models
- After that apply models for prediction
- It's necessary to use some statistical validation techniques to determine the accuracy.

The Accuracy result has been successfully displayed.

A. Data Set

We made use of a Kaggle dataset that was specially picked for its capacity to enable us to foretell the presence or absence of diabetes in a patient based on diagnostic readings. Despite facing numerous limitations during the selection process, we were able to extract a subset of the larger dataset that was suitable for our purposes. The problem at hand is a classic example of supervised binary classification, where we aim to categorize patients as either positive or negative for diabetes. In this study used the dataset contains 768 records describing female patients of the Pima Indian community, with 9 attributes, including a class attribute. We identified a distribution of 268 positive instances (34.9%) and 500 negative instances (65.1%) across these entries. Table 1 provides a thorough discussion of each feature.

TABLE 1. DESCRIPTION AND CHARACTERISTICS OF DATASET

Serial	Name of Attribute	Description of Attribute
I	Glucose	Glucose levels during a 120-minute test for oral glucose tolerance
II	Pregnancies	Number of pregnancies a woman had
III	BP	Diastolic Blood Pressure
IV	Insulin	2-hour serum insulin
V	Skin Thickness	Skin's thickness in folds
VI	Diabetes Pedigree	Diabetes pedigree Function
VII	BMI	Weight/(height)2 = BMI
VIII	Age	Age (Years)
IX	Outcome	For diabetes, the class variable has a value of '1' for Positive and '0' for Negative.

B. Processing the Data

Real-world data can be difficult to work with because of a variety of issues, including missing numbers, erratic and inaccurate information, and poor entire data quality. Without proper preprocessing, obtaining quality results from such data can be nearly impossible. Data preparation methods including cleaning, integrating, transforming, reducing, and discretizing are used to address these problems. While considering variables such as time, cost, and quality, these techniques seek to make more data suitable for mining and analysis. We can apply appropriate preprocessing strategies to improve data quality, and increase the likelihood of discovering valuable insights and knowledge from it.

C. Data Cleaning

An essential part of data preprocessing is data cleaning that involves identifying and resolving issues such as missing values and noisy data. Missing values can be filled in using imputation techniques,

while noisy data can be detected and resolved by removing outliers. We found that certain characteristics in the dataset we utilized for our analysis, including glucose, skin thickness, BP, BMI, and insulin, had zero (0) values. In order to tackle this, we took the standard imputation strategy and substituted all zero values with the attribute's median value. By performing data cleaning and imputation in this manner, we were able to ensure that our dataset was ready for further analysis and modeling.

D. Random Forest

To predict early diabetes using random forest, we need to first collect a dataset of relevant features and corresponding labels. The features could include demographic information, lifestyle factors, and biomarkers has a relation to diabetes, such as insulin resistance, Blood Glucose levels, and Body Mass Index (BMI). Once we have our dataset, we use a random forest algorithm to train a model to predict the presence of early diabetes based on the selected features. During training, the algorithm will randomly select subsets of the data and features to create multiple decision trees. Each decision tree will make a prediction, and the results will be combined to produce a final prediction. We measure our model's performance using measures like recall, accuracy, and precision. In order to make sure that our model can be applied to fresh data, we also employ strategies like cross-validation. By utilizing numerous decision trees trained on subsets of the available data, random forest is a robust and adaptable technique that may be used to detect early diabetes. However, the quality and relevance of the features used to train the model will be critical in achieving accurate and reliable predictions.

E. Naive Bayes

Naive established on the Bayes theorem, the Bayes algorithm is a probabilistic Machine Learning method. It is a multi-class classification algorithm that may be applied to issues with binary classification. Assuming that the characteristics in the input data are free from each other which is generally not the case in current-world application, is what gives the algorithm its "naive" name. Despite this oversimplification, the technique may still function effectively and is often used in a variety of applications, including document categorization, sentiment analysis, and spam filtering.

Using Bayes' theorem, the algorithm determines the conditional probability of every class given the input characteristics, which states: Q(b|a) = Q(a|b) * Q(b) / Q(a).

Where Q(b|a) denotes the likelihood of class b given a given input feature, Q(a|b) denotes the likelihood of observing input features of a given class b, Q(b) denotes the prior likelihood of class b, and Q(a) denotes the likelihood of witnessing input characteristics a.

F. Support Vector Machine

To predict the presence of diabetes SVMs can be used. The process involves collecting a dataset such as Body Mass Index (BMI), age, glucose levels, and BP with relevant features, preparing the data, selecting the most relevant features, training an SVM model using the selected features, evaluating the model's performance, fine-tuning the model if necessary, and using it to make predictions on new data. The most popular supervised classifier is SVM, which uses a hyperplane to categories data in N-dimensional space. [9]. The predictions will indicate the likelihood of a patient having early diabetes.

G. Decision Tree

The visual representation of a decision-making process in a tree format is called a decision tree. Each leaf node represents a choice or potential course of action, whereas every inside node indicates a test on an attribute, its result, and its branch. Together, these three nodes form a tree-like structure. It provides high accuracy and stability in the form of a tree [10]. Decision trees are often used in machine learning and artificial intelligence applications for classification and prediction problems.

F. Boosting Algorithm

Boosting is a form of machine learning method that makes weak models stronger. The fundamental idea behind boosting is to train many weak models in a

cycle on various subsets of training data, with each new model being trained to rectify the mistakes of the preceding models. Although there are many different boosting algorithms, AdaBoost (Adaptive Boosting) is one of the most well-known. AdaBoost works by assigning weights to each training example, with weights initially set to equal values. First, the weak model is trained on the data and then its accuracy is evaluated. The weights are then adjusted such that examples that were misclassified by the first model have higher weights, and examples that were correctly classified have lower weights. A second weak model is then trained on the adjusted data, and this process is repeated iteratively for a fixed number of rounds or until the accuracy of the model reaches a predetermined threshold.

IV. Result and Discussion

Better health outcomes and the prevention of complications can result from the beginning detection and management of diabetes. A person's various characteristics including age, blood pressure, skin thickness, insulin, and family history of the illness, ML algorithms can be used to forecast chance of acquiring diabetes.

To predict diabetes, we employ a variety of ML algorithms, such as Decision Tree, RF, Naive Bayes, Support Vector Machine and Boosting Algorithm. When the total sample size is 768, the total training size is 460 and the total test size is 308, we get an accuracy of 0.75 for Random Forest, 0.7597 for Naive Bayes, 0.75 for Support Vector Machine, 0.68 for Decision Tree and 0.75 for Boosting algorithm.

Confusion matrix

	P	N
T	184	21
F	53	50

Accuracy Score	0.7597402597402597

Overall, machine learning algorithms show promise in predicting the risk of diabetes and its complications. However, it is important to note that machine learning algorithms should not replace clinical judgment and diagnosis by healthcare professionals. Healthcare workers should employ machine learning algorithms as an additional tool to help them make wise decisions about patient care.

V. Conclusion

In conclusion, the early detection and segmentation of diabetes based on machine learning approaches have shown promising results. The model gave 75.97% accuracy score. The use of machine learning models has made it possible to predict and categorize diabetes with a high degree of efficiency. The utilization of feature extraction techniques and classification algorithms has shown significant improvement in the accuracy of diabetes detection and segmentation compared to traditional methods. With the aid of machine learning algorithms, clinicians and medical professionals can now detect and segment diabetes at an earlier stage, which can result in to better treatment and management of the disease.

Moreover, machine learning algorithms have the potential to overcome the limitations of conventional diabetes diagnosis techniques, such as the dependence on expert knowledge and subjectivity. These algorithms can learn from the data and improve their accuracy with each new input. However, despite the significant advancements in the field of machine learning and diabetes detection, there are still some challenges that need to be addressed. For instance, the data collection process needs to be improved to ensure that the dataset is representative and diverse. Additionally, the interpretability of machine learning models needs to be enhanced to ensure that they are clinically relevant and can be easily understood by medical professionals.

In conclusion, the early detection and segmentation of diabetes based on machine learning approaches have shown great potential in revolutionizing the diagnosis and management of diabetes. Further research and development in this field can lead to a better understanding, treatment, and management of diabetes, in the end, improving millions of people's standard of living around the world.

References

1. Saru S, Subashree S. Analysis and prediction of diabetes using machine learning. International journal of emerging technology and innovative engineering. 2019 Apr 2;5(4).
2. Aada A, Tiwari S. Predicting diabetes in medical datasets using machine learning techniques. Int. J. Sci. Res. Eng. Trends. 2019;5(2):257-67.
3. Alehegn M, Joshi R, Mulay P. Analysis and prediction of diabetes mellitus using machine learning algorithm. International Journal of Pure and Applied Mathematics. 2018;118(9):871-8.
4. Sisodia D, Sisodia DS. Prediction of diabetes using classification algorithms. Procedia computer science. 2018 Jan 1;132:1578-85.
5. Mujumdar A, Vaidehi V. Diabetes prediction using machine learning algorithms. Procedia Computer Science. 2019 Jan 1;165:292-9.
6. Krishnamoorthi R, Joshi S, Almarzouki HZ, Shukla PK, Rizwan A, Kalpana C, Tiwari B. A novel diabetes healthcare disease prediction framework using machine learning techniques. Journal of Healthcare Engineering. 2022 Jan 11;2022.
7. Zou Q, Qu K, Luo Y, Yin D, Ju Y, Tang H. Predicting diabetes mellitus with machine learning techniques. Frontiers in genetics. 2018 Nov 6;9:515.
8. Islam S.M.S., Hasan M.M. and Abdullah S., 2018. Deep learning based early detection and grading of diabetic retinopathy using retinal fundus images. *arXiv preprint arXiv:1812.10595*. (2021).
9. Revathi A, Kaladevi R, Ramana K, Jhaveri RH, Rudra Kumar M, Sankara Prasanna Kumar M. Early detection of cognitive decline using machine learning algorithm and cognitive ability test. Security and Communication Networks. 2022 Jan 20;2022:1-3.
10. Tigga NP, Garg S. Prediction of type 2 diabetes using machine learning classification methods. Procedia Computer Science. 2020 Jan 1;167:706-16.