Diabetes Detection Using Machine Learning Algorithms

Nicole D'Souza
MPSTME, NMIMS
Mumbai,India
nicolemichelledsouza@gmail.com

Kunjal Shah
MPSTME, NMIMS
Mumbai,India
kunjal2001@hotmail.com

Pranav Singh
MPSTME,NMIMS
Mumbai,India
singh.pranav161@gmail.com

***Abstract*—Diabetes is a serious illness. Predicting this disease in a timely manner is necessary to avoid severe side effects. Current medical practise dictates that a patient undergoes a battery of tests in order to obtain the information necessary for diagnosis, after which treatment is administered based on the diagnosis. However, in many cases, the early stages go undetected, and it is quite difficult for physicians to diagnose due to the interdependence of numerous factors. A single parameter is commonly inadequate for the accurate diagnosis of diabetes and may lead to erroneous decisions. To accurately forecast diabetes at an early stage, multiple criteria must be combined. This study proposes the development of an early diabetes detection model. The model will not only be more accurate than humans, but it will also reduce the workload of medical professionals.**

***Index Terms*—machine learning, diabetes detection, ensemble learning**

I. Introduction

Diabetes Mellitus is a pressing public health problem. It is caused by several factors, including heredity, obesity, insulin resistance, physical inactivity, and many more. As of 2014, approximately 422 million people worldwide were diagnosed with diabetes (World Health Organization [WHO], 2020), and it is estimated that 700 million adults shall be afflicted by 2045. (International Diabetes Federation[IDF],2020). To avoid complications, a growing number of researchers are focusing on developing a method for accurately predicting diabetes at an early stage. Early detection of diabetes mitigates the risk of developing conditions like renal failure, heart attacks, stroke, blindness, and amputation of the lower extremities. [1].

Because of the significant social impact of this particular condition, a massive volume of data is unavoidably generated. Thus, data mining and machine learning are effective methods for extracting usable information from vast databases. Various techniques have been developed within this study, and an ensemble classification strategy employing data mining and machine learning has been proposed for diabetes.

Data mining is the study of strategies for gaining knowledge from a database. Intriguing patterns in massive datasets are identified and analysed. Data mining entails clustering, association, and categorization of data. Various categorization approaches are available, such as Naive Bayes, k-nearest neighbours, decision trees, regression, and artificial neural networks. These classification techniques have been implemented in medical, commercial, and industrial settings. Numerous researchers utilised the Pima Indian Diabetes Dataset for diagnosis. There are eight parameters. This includes the body mass index,diastolic and systolic blood pressure, plasma glucose, skinfold thickness, diabetic pedigree function, number of pregnancies, and Type 1 or Type 0 diabetes (where 0 stands for non-diabetic and 1 means diabetic patient) [2].

II. Related Works

In a study by [3], a data set of 1004 samples with nine attributes was utilised. The primary purpose was to develop a model based on an ANN to identify the effective variables and their diverse effects on diabetes. The various types of factors were classified as input and output variables. The set's data was input into the JNN tool environment, which was trained, validated, and tested to achieve an accuracy of 87.3%.

In a paper written by [4], a real-time dataset containing various risk factors of diabetes mellitus was collected from approximately 200 individuals. The study conducted a comparative analysis of various ML techniques, and based on the evaluation, applicable risk factors are predicted. The results demonstrated that the C4.5% decision tree gave the most accurate data classification.

A procedure for estimating gestational diabetes using Microsoft Azure AI services was outlined in [5]. UCI's PIMA Indian dataset was utilised. This dataset's predictive model was constructed and cross-validated. In the initial stages of pregnancy, the classification model predicts the occurrence of gestational diabetes based on a variety of factors. After analysing the algorithm and collecting 768 samples, the highest accuracy of 77.8 percent was attained.

In another recent study done by [6] deep learning method was used to create a huge diabetes forecast system. The goal was to lower the amount of false positives and negatives to the maximum and reach the highest precision. For diabetes prediction ELM classifier was applied as its ability to learn is quick. Out of all the four suggested classifiers DL had the highest percentage of accuracy of 98.07%.

In a study model done by [7] predicting type II diabetes based on the random forest method was aimed at analyzing readily available indicators. The random forest comprises multiple decision trees as an ensemble classifier, it has high

978-1-6654-9291-1/22/$31.00 ©2022 IEEE

accuracy and is robust. Dataset used was from the University of Virginia. The authors intend to build on more indications for predicting the risk of diabetes, as well as update the future viewpoint of data mining.

In a different study by [8], the selection of features to identify diabetes mellitus and the design of a predictive algorithm that would use those features to determine the optimal classifier that would provide the closest results to clinical outcomes were analysed. Utilizing predictive analysis, the strategy employed a subset of failed Diabetes Mellitus early diagnosis features. The decision tree algorithm and the random forest algorithm achieved the highest accuracy of 98.20 percent and 98.00 percent, respectively, and are therefore the most effective techniques for analysing diabetic data. At 82.30 percent, the Naive Bayesian has the most accuracy. In addition, the study generalises the selection of significant factors and characteristics from the dataset in order to improve classification precision.

In another work conducted by [9] utilising the Waikato Environment for Knowledge Analysis toolbox, the authors proposed a type 2 diabetes prediction model based on data mining techniques. The aims were to improve the accuracy of the prediction model and enable the model to adapt to multiple datasets. The model has two components, a modified K-means method and logistic regression technique, after various preprocessing steps. Upon comparing results, it was determined that the model was 3.04 percent more accurate than those of other studies. In addition, their model confirmed that the datasets quality was adequate. To further examine the efficacy of the two models, they were applied to two more diabetes datasets that produced favourable outcomes.

In [10], The authors recommended employing deep neural networks to diagnose diabetes by training its components with five and tenfold cross-validation. The Pima Indian Diabetes (PID) dataset was taken from UCI's machine learning repository database. The findings of five-fold cross-validation indicated that the deep learning strategy has a prediction accuracy of 98.35 percent, which is higher than the other approaches employed.

A paper written by [11] intended to implement a prediction model for measuring diabetes risk. On the PIMA dataset, different classifiers were applied, demonstrating the competency of data mining and machine learning algorithms that reduce risk factors and produce high efficiency and accuracy. The accuracy of the four classifiers (DT, ANN, NB, and DL) ranged from 90% to 98%, which was significantly higher than the accuracy of the available algorithms. About 98.07 percent of the time, DL is the most accurate of these four approaches. The authors intend to improve the robustness of the system by creating an app or website that use the DL algorithm to assist healthcare practitioners in the early detection of diabetes. In another study by [12], The authors sought to construct a model capable of properly predicting the risk of diabetes in patients. The three ML classification techniques employed were decision tree, SVM, and Nave Bayes. The dataset utilised was the Pima Indians Diabetes Dataset from the UCI machine learning library. The performance was then classified according to its recall, F-measure, and precision. In comparison to other algorithms, the results indicated that the Nave Bayes method had the highest precision at 76.30 percent. ROC (Receiver Operating Characteristics) curves were then utilised to systematically confirm the accuracy of these data. In a study done by [13] , the PIDD was taken and applied to five different data mining algorithms. These five algorithms and their accuracies are a Gaussian mixture model which gave 81.5 accuracy, Artificial Neural Networks with the accuracy of 89%, ELM with an accuracy of 82%, the accuracy of Logistic Regression and Support Vector Machine is 64% and 74%, respectively. The highest accuracy was found in ANN. In another paper by [14] the authors implemented several algorithms including a few ensemble learning algorithms as well. The most accuracy was obtained by Logistic regression at 96% followed by LDA 94%. In a paper by [15] diabetes is predicted and the link between the various traits is also defined. Diabetes attribute selection, grouping, prediction, and association rule mining are determined using a variety of techniques. Significant attributes selection was done by a method known as component analysis. The Apriori method indicates a significant association between diabetes, body mass index (BMI), and glucose level. For diabetes prediction, ANN, RF, and K-means clustering algorithms were employed. ANN generated the most accurate predictions with 75.7%. In another research conducted by [16] the authors aimed to develop a model that could accurately predict the risk of diabetes in patients. Decision tree, SVM, and Nave Bayes were the three ML classification approaches utilised. The dataset utilised was the UCI machine learning library's Pima Indians Diabetes Dataset. The performance was subsequently categorised based on its F-measure, recall, and precision. The results indicated that, in comparison to other algorithms, the Nave Bayes technique had the highest precision at 76.30 percent. ROC (Receiver Operating Characteristics) curves were then utilised to confirm the accuracy of these data in a systematic manner. The study done by [17] For accurate authentication, numerous algorithms, including Decision Tree, Support Vector Machines, Nave Bayes, and ANN, were implemented. Using SVM and ANN, the greatest accuracy of 82% was attained.

A new methodology is proposed by [18] That is, using PCA to alter the initial collection of features, thereby resolving the correlation problem that makes it difficult for classification algorithms to discover correlations among the data. The PCA will aid in filtering out unnecessary data, such as irrelevant features, hence decreasing training time, cost, and enhancing model performance. Due of k-means' capacity to account for outliers in the data, unsupervised K-means clustering is used following PCA analysis. The K-means clustering result is further sanitised, and Logistic regression is applied to the dataset to create a supervised classification model. The model's 89% accuracy can be improved by the use of real-time data.

III. RESEARCH GAPS

This section aims to identify the gaps in our study and what can be done to solve them in the future.

The first obstacle relates to data collection. [25], [26] When performing a literature review, finding articles and publications unrelated to the PIMA India dataset was one of the greatest obstacles. Other datasets are either very tiny, insufficient, or lack real-time data, in contrast to this one, which is well-established and yields somewhat satisfactory findings. Even though tiny datasets demonstrate great accuracy, this is due to the model's overfitting, and these models perform poorly on additional testing data. Thus, implementation in real time is not practical. To maintain the training and optimization of models, it is vital to incorporate current and accurate patient data. The size of the dataset should be sufficient to correctly train the model, enabling it to make more accurate predictions.

A second issue is feature selection. Some authors overlooked a few essential elements, while others bundled them for training convenience and efficacy. It is vital to choose a model with performance-optimizing features [**?**].

Debugging was also problematic because tools that divide code into cells, such as Jupyter Notebooks, cannot be utilised on automation batch processes. A number of authors required a time-series data set. It is difficult to duplicate this work because no internet resource contains it. Such sophisticated models require intensive tuning and substantial data sets for training and testing both.

Another difficulty is the creation of a physical model. To obtain perfect accuracy, numerous parameters must be changed. While constructing a model, variables such as kernel, random states, number of trees, and hyper parameter tuning are considered, among others. Choosing the suitable algorithm with the appropriate hyperparameters is crucial. Some classification models' real-time detection accuracy is diminished since they only train on one parameter. A comprehensive review of these schemes reveals that the bulk of them either have a single input parameter or feature selection that is suboptimal.

Furthermore, some classification-based systems are dependent entirely on particular hardware components, making their availability and flexibility even more challenging.

IV. METHODOLOGY

Figure 1 shows the methodology followed. We used a diabetes dataset with 2000 instances and 9 attributes to train the model [20], then displayed the data and conducted feature scaling with a Standard Scaler. EDA was carried out to better comprehend the data we obtained. Following that, we pre-processed our data, which means we had to clean it by removing duplicate, missing, or strange values and scaling it properly. The next stage was to choose the machine learning algorithms that would train the data. The model's accuracy was then tested using the test dataset. Finally, the models were compared based on several performance indicators such as accuracy, f1-score, and recall among others

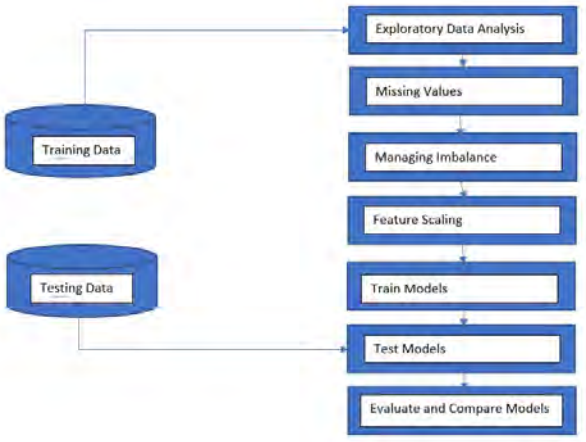

Fig. 1. Methodology

A. Dataset Attributes

The diabetes dataset used for this project contained 2000 instances and 9 attributes. A description of those values is given below. The data was split into 80% for training and 20% for testing.

TABLE I
DATASET DETAILS

Attribute	***Range***	***Description***
Pregnancies	0-17	Number of pregnancies
Glucose	0-199	Plasma Glucose Concentration
Blood Pressure	0-122	Diastolic Blood Pressure
Skin Thickness	0-110	Triceps skin fold thickness
Insulin	0-744	2-hour serum Insulin
BMI	0-80	Body Mass Index
Pedigree Function	0.078-2.42	Possibility based on family history
Age	21-81	Age (years)
Outcome	0 or 1	0 = healthy,1= diagnosed diabetes

B. Data Preprocessing

Machine Learning Algorithms rely heavily on the data used to train them, real life data is often times imbalanced, it may contain a lot of noise, missing values and outliers this affects the quality of the model produced [21]. In the data preprocessing phase, issues of imbalance and missing values are addressed, and the dataset is split 80:20 for training.

1) **Missing Values** :
   In large datasets missing values can be managed we simply deleting them, however in this case since the dataset was only 2000 rows, we made use of a KNN imputer. A KNN imputer imputes the missing value on the basis of the value of the K nearest neighbour [22].
2) **Managing Imbalance** :
   Imbalance is commonly occurring problem in machine learning where the number of records in both classes is not equal. An imbalanced dataset causes the model to be biased and due to this bias, it often times fails to

appropriately predict the minority class [23] . To address this issue, we employ the random over sampler, that manages imbalance by randomly replicating rows from the minority class.

3) **Feature Scaling** :
Feature scaling is carried out to normalize the range of independent variables, this is commonly referred to as data normalization. In [24], the feature scaling technique used is Standard Scaler. A standard scaler resizes the distribution of the values so that each feature has the mean =0 and standard deviation =1.

C. Algorithms Used

1) **K Nearest Neighbours** :
It is a Supervised Learning regression and classification technique. It is a technique frequently used to resample datasets. According to its name, it assigns a class to an unknown data point based on the K nearest known data points.
2) **Logistic Regression** :
The Logistic Regression Classifier was used as the initial model. It calculates a weighted sum of the input characteristics, but instead of determining the outcome, it calculates the probability of an outcome based on the individual factors.
3) **Decision Tree** :
A tree or graph-like structure is built using characteristics such as cost, categorization categories, and effort. The selection is made by moving from root to leaf until all of the requirements are met. Gini indices are used to determine node split. The incorporation of Gini indexes aids in node splitting.
4) **Naive Bayes** :
Naive Bayes is based on the Bayes theorem. It is a type of algorithm in which the value of one property is assumed to be independent (naive) of the value of another feature. It considers conditional probability, which determines the likelihood of an event occurring if some of the other occurrences have already occurred.
5) **Random Forest** :
The Random Forest algorithm is a versatile classification and regression algorithms as it merges multiple decision trees into one entity. Each tree further classifies the data according to the class to which it belongs, and the final class is the one with the most votes.
6) **Gradient Boosting** :
A number of weak learners' predictions are combined by the Gradient Boosting Machine (GBM). Each node of a decision tree employs a unique set of attributes to determine the optimal split. In addition, each successive tree takes into account the errors or flaws committed by the trees that came before it. Therefore, each successive decision tree is constructed based on the previous trees' errors. In this manner, successive trees are generated for the gradient boosting machine method.
7) **XGB** :
XGBoost is nothing more than an improved version of the GBM algorithm! XGBoost builds trees consecutively, attempting to repair the mistakes of earlier trees. A fundamental difference is that XGBM employees parallel pre-processing, which allows it to be quicker than GBM. In addition, XGBoost incorporates a number of regularization algorithms that prevent overfitting and increase overall performance.
8) **LGBM**: Because of its speed and efficiency, the LightGBM boosting algorithm is becoming increasingly popularThe trees in LightGBM grow leaf-by-leaf rather than level-by-level. Following the initial split, the next is performed exclusively on the leaf node with the highest delta loss.

V. RESULTS

Table 2 shows the results obtained. The Random Forest algorithm demonstrated the most accuracy of 98.1%, followed by Decision Tree at 98%, and XGB and LGBM at 97.6%. The remaining models obtained an accuracy between 77-88%.

TABLE II
METRICS FOR TESTING DATA

Algorithm	***Accuracy***	***Precision***	***Recall***
K Nearest Neighbours	88.2%	97%	87%
Logistic Regression	77.2%	90%	79%
Decision Tree	98%	97%	100%
Naive Bayes	75.8%	85%	80%
Random Forest	98.1%	97%	98%
XGB	97.6%	96%	99%
LGBM	97.6%	96%	99%
Gradient Boosting	83.44%	81%	86%

VI. CONCLUSION & FUTURE SCOPE

Various ML algorithms were utilised in this study, and it was determined that Random Forest provides the most classification accuracy at 98.1%. We compared the outcomes of multiple ML algorithms on the data set to demonstrate that the model provides a more accurate and precise prediction of diabetes than the existing data set.

This work can be expanded to estimate the number of non-diabetic people who will develop diabetes in the coming years. In the future, we must also develop feature selection techniques such as SVD, PCA and LDA to better the accuracy of the models and enhance their performance. This will allow us to better select important features and consequently reduce the model's training time.

Implementing Neural Networks such as ANN, CNN, and RNN in conjunction with other algorithms will aid in the detection of diabetes with greater precision, because hybrid schemes are necessary for improving the performance of models.

This model was trained with the help of a structured dataset. However, unstructured data must also be taken into consideration in the future, as this will provide a clearer picture

of the model's accuracy in the lack of a vast amount of structured data.

Additionally, attributes such as family history and lifestyle habits can be included in the model.

Due to early detection, patients can be treated more effectively, thereby preventing further risks and reducing the likelihood of future dangers. The insights gained from these diabetes detection models can be applied to various other medical domains for prediction, like detection of sleep apnea, various types of cancer, patterns of mental health disorders, etc.

REFERENCES

[1] World Health Organization, "Diabetes," 16 September 2022. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/diabetes.

[2] V. Chang, J. Bailey, Q. A. Xu, and Z. Sun, "Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms," Neural Computing and Applications, 2022.

[3] Nesreen Samer El-Jerjawi, "Diabetes Prediction Using Artificial Neural Network," International Journal of Advanced Science and Technology, vol. 121, pp. 55-64, 2018.

[4] M. F. Faruque, Asaduzzaman, and I. H. Sarker, "Performance analysis of machine learning techniques to predict diabetes mellitus," 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), 2019.

[5] Y. Srivastava, P. Khanna, and S. Kumar, "Estimation of gestational diabetes mellitus using azure AI services," 2019 Amity International Conference on Artificial Intelligence (AICAI), 2019.

[6] R. Bhargava and J. Dinesh, "Deep Learning based system design for diabetes prediction," 2021 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON), 2021.

[7] Z. Xu and Z. Wang, "A risk prediction model for type 2 diabetes based on weighted feature selection of Random Forest and XGBoost ensemble classifier," 2019 Eleventh International Conference on Advanced Computational Intelligence (ICACI), 2019.

[8] N. Sneha and T. Gangil, "Analysis of diabetes mellitus for early prediction using optimal features selection," Journal of Big Data, vol. 6, no. 1, 2019.

[9] H. Wu, S. Yang, Z. Huang, J. He, and X. Wang, "Type 2 diabetes mellitus prediction model based on data mining," Informatics in Medicine Unlocked, vol. 10, pp. 100–107, 2018.

[10] S. Islam Ayon and M. Milon Islam, "Diabetes prediction: A deep learning approach," International Journal of Information Engineering and Electronic Business, vol. 11, no. 2, pp. 21–27, 2019.

[11] H. Naz and S. Ahuja, "Deep Learning Approach for diabetes prediction using Pima Indian Dataset," Journal of Diabetes; Metabolic Disorders, vol. 19, no. 1, pp. 391–403, 2020.

[12] D. Sisodia and D. S. Sisodia, "Prediction of diabetes using classification algorithms," Procedia Computer Science, vol. 132, pp. 1578–1585, 2018.

[13] M. Komi, Jun Li, Yongxin Zhai, and Xianguo Zhang, "Application of data mining methods in diabetes prediction," 2017 2nd International Conference on Image, Vision and Computing (ICIVC), 2017.

[14] A. Mujumdar and V. Vaidehi, "Diabetes prediction using machine learning algorithms," Procedia Computer Science, vol. 165, pp. 292–299, 2019.

[15] T. Mahboob Alam, M. A. Iqbal, Y. Ali, A. Wahab, S. Ijaz, T. Imtiaz Baig, A. Hussain, M. A. Malik, M. M. Raza, S. Ibrar, and Z. Abbas, "A model for early prediction of diabetes," Informatics in Medicine Unlocked, vol. 16, p. 100204, 2019.

[16] M. A. Sarwar, N. Kamal, W. Hamid, and M. A. Shah, "Prediction of diabetes using machine learning algorithms in Healthcare," 2018 24th International Conference on Automation and Computing (ICAC), 2018.

[17] Sonar, Priyanka, and K. JayaMalini, "Diabetes prediction using different machine learning approaches," in 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), 2019.

[18] Zhu, C., Idemudia, C. U., Feng, W., "Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques," Informatics in Medicine Unlocked, vol. 17, 2019.

[19] Kumari, S., Kumar, D., Mittal, M, "An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier," International Journal of Cognitive Computing in Engineering, vol. 2, pp. 40-46, 2021.

[20] M. Lichman, "Pima Indians diabetes database," [Online]. Available: https://archive.ics.uci.edu/ml/datasets/diabetes

[21] J. Malley B, "Data Pre-processing.," in MIT Critical Data, editor. Secondary Analysis of Electronic Health Records, Springer, 2016.

[22] L Beretta, A. Santaniello, "Nearest neighbor imputation algorithms: a critical evaluation," Selected articles from the 5th Translational Bioinformatics Conference (TBC 2015): medical informatics and decision making, 25 July 2016.

[23] "A review on imbalanced data handling using undersampling and oversampling technique," International Journal of Recent Trends in Engineering and Research, vol. 3, no. 4, pp. 444–449, 2017.

[24] V. N. Raju, K. P. Lakshmi, V. M. Jain, A. Kalidindi, and V. Padma, "Study the influence of normalization/transformation process on the accuracy of supervised classification," 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), 2020.

[25] Sheikhi G, Altınçay H (2016) The cost of type II diabetes mellitus: a machine learning perspective. In: Kyriacou E, Christofides S, Pattichis CS (eds) XIV mediterranean conference on medical and biological engineering and computing 2016. IFMBE proceedings, vol 57. Springer, Cham, pp 818-821.

[26] Iyer A, Jeyalatha S, Sumbaly R (2015) Diagnosis of diabetes using classification mining techniques. Int J Data Min Knowl Manag Process 5(1): 1–14. https://doi.org/10.5121/ijdkp.2015.5101

[27] Sharma, T., Shah, M. A comprehensive review of machine learning techniques on diabetes detection. Vis. Comput. Ind. Biomed. Art 4, 30 (2021).