A Survey on Diabetes Detection Based on MachineLearning Classifiers

Pratham Thakral
Research Scholar, Chandigarh University, Punjab
pthakral1998@gmail.com

Jasminder Kaur Sandhu
Assistant Professor, Chandigarh University, Punjab
jasminder.e12840@cumail.in

2023 International Conference on Circuit Power and Computing Technologies (ICCPCT) | 979-8-3503-3324-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICCPCT58313.2023.10245522

Abstract—Computer-assisted sickness diagnosis is less costly, time-saving, accurate and it eliminates the need for extra employees in medical decision-making. According to many nutrition studies, nearly a quarter of the world's population suffers from chronic illnesses such as diabetes. As a result, an effective ma- chine learning regressor capable of correctly diagnosing diabetes is urgently needed. The aim is to determine whether individual classifiers or a set of classifier combinations provide the best diabetes detection accuracy. The concentration of glucose in human blood is known as the blood glucose level. Diabetes is a long-term condition marked by a lack of glucose. To maintain blood sugar levels within the suggested target range, insulin therapy is necessary. As per the statistic from the WHO, diabetes affects almost 400 million individuals worldwide. The need for regular blood glucose monitoring in the therapy phase cannot be overstated. Diabetes can be detected in a variety of ways. ML models are utilized to determine whether or not the patient has diabetes to confirm the sickness. This paper attempts to survey various recently published literature that have proposed different approaches to detecting Diabetes Mellitus using machine learning techniques. Some of the surveyed techniques utilise comparative analysis of ML as well as DL classifiers, while others techniques use medical records to identify the risk predictor.

Index Terms—Random Forest (RF), Machine Learning (ML), Deep Learning (DL), Logistic Regression (LR), Decision Tree (DT), Support Vector Machine (SVM), Gaussian Naïve Bayes (NB), K-Nearest Neighbour (KNN).

I. INTRODUCTION

One of the most critical problems now facing the medical sector is diabetes, and its impact is expanding quickly. As a result, WHO ranked it as the world's sixth largest cause of premature death in 2016 [1]. Each year, 1.6 million patients died as a consequence of diabetes throughout the world [2]. As per the first global WHO report, the number had increased to 422 million (8.5%) by the end of 2014 from 108 million (4.2%) [3].WHO joined partners from all across the globe to highlight the effects of diabetes on World Diabetes Day 2018. According to the WHO, one in three individuals has overweight, as well as the circumstance is becoming deteriorating. Diabetes has been linked to heart attacks, renal failure, and stroke-related blindness [1]. A condition known as diabetes is characterized by excessively elevated blood glucose levels. Diabetes develops when the pancreatic gland in the human body is unable to produce enough insulin, according to medical experts. DM1 requires insulin, as well as the insulin generated, can't be used via cells (Type 2 diabetes) [4]. Diabetes mellitus: This chronic disease, sometimes called "diabetes," is characterized by extremely elevated blood sugar and glucose levels. Either a lack of insulin synthesis or tolerance in the cells causes diabetes. Diabetes mellitus can also occur as a side effect of another ailment like myotonic dystrophy, pancreatic illness, or drugs like glucocorticoids. Diabetes that develops during pregnancy is known as gestational diabetes. According to the National Diabetes published in 2011, around 25.8 million Americans nearly 8.3% of the population is facing diabetes. Pre-diabetes has also been detected in around 79 million persons [5]. A significant number of data is created in big data and ML has evolved into an essential technique for analyzing data's generated complexity. Single classifier and classifier ensemble

In article [6], two strategies have been used for data analysis in medical diagnosis. With machine learning models, health information effectiveness can be increased, patient counts can be lowered, as well as healthcare expenses might also be cut. As a consequence, these models are commonly used to conduct medical analysis while comparing to alternative traditional approaches. The only method to reduce overall rates of death affected by Chronic Diseases (CDs) is to identify them soon as well as cure them effectively. As an outcome, the majority of health experts are intrigued by new forecasting modeling methods for sickness forecasting [7]. In Diabetes Mellitus (DM) research, using ML and Data Mining approaches is a fundamental strategy for extracting knowledge from enormous amounts of available diabetes-related data. Because of the disease's significant socio-economic effect, it is few of its top priorities in healthcare study, which naturally forms massive volumes of information [8]. Data mining is a huge step forward even in fields of analytical techniques. It has been demonstrated an incorporating data mining into medical analysis improves diagnosis accuracy, lowers expenses, and saves human capital [9].

The purpose of this investigation is to investigate the outcomes of the machine learning (ML) algorithms that have demonstrated the greatest success in predicting the onset of diabetes. Because we have access to the performance parameters of many machine learning algorithms, which were gathered during an experiment on the same input (i.e., the same dataset), we have been able to draw inferences about these algorithms. These parameters were collected during an experiment on the same input. This not only highlights the value of ML in the healthcare industry, but also demonstrates its capacity to provide correct predictions and assist practitioners. The following is the order in which the remaining portions of the paper are presented: A literature review that gives an overview of the projected future research activity in this topic may be found in the "Section II" portion of this report. In the third section, we will explore the preliminary work as well as the principles of machine learning. Classification techniques as well as the Pima Indians Diabetes Dataset (PIDD), which is the dataset that is utilised in academic circles the most frequently. In the following section (IV), the experimental analysis and results pertaining to PIDD are discussed. The conclusion and any consequences it may

979-8-3503-3324-4/23/$31.00 ©2023 IEEE

have for the future are discussed in the fifth and last portion of the paper, titled portion V.

II. LITERATURE REVIEW

The Pima Indian diabetes dataset was utilized in several studies to predict diabetes. 9 variables, 768 entries characterizing female patients in PIDD. This section discusses several works that are closely related [10]–[34].

Ahmed et al. [10]'s When applied to real-time medical data, the fusion machine learning method resulted in an accuracy of 94.87 percent, outperforming both the SVM and ANN algorithms. However, keep in mind that the technique achieved a higher AUC score than its competitors by a margin of 12.16%.

A. Banerjee and N. Deepa [11] The images of the tongue were used as input in the development of a computer-aided model that provides intelligent decision support. The identification and categorization of diabetes were both accomplished with the help of this approach. This model captures every detail, including the spots and texture of the fur, as well as the color and coating of the fur, and even the dental markings that are found on the tongue. Following that, the data that were generated are classified using SVM, then PSO, and lastly confirmed by a real-time dataset in order to evaluate, contrast, and demonstrate the respective efficacy and efficiency of each method. The utilization of images was crucial in achieving the accuracy of 97.82 percent.

Fazakis and others [12] After developing models for predicting the risk of type-2 diabetes using Naive Bayes (NB), DL, random forests, artificial neural networks (ANN), and deep neural networks, the researchers constructed an ensemble learning model that contained the following components: Weighted Voting System, LR-RF Voting System AUC values of were found in the EL models that they developed. AUC values were obtained from their models that were based on stacking, voting, and weighted voting respectively. 83.3%, 88.1%, and 88.4%, sensitivity values of 77.3%, 79.4%, and 85.6%, and specificity values of 88.4%, 79.2%, 84%, and 79.8%, respectively.

Sivaranjani et.al. [13]'s A comparison of SVM, NB, LR, and RF (without pre-processing) as well as RF (with pre-processing) is suggested as future work. The most accurate results were obtained using ML algorithms i.e., 83 percent.

After using the PID and NHNES datasets, Syed and Khan [14] designed and in order to determine the likelihood of getting type 2 diabetes, we analyzed and screened the most significant diabetes uncertainty condition using Chi-squared analysis and binary logistic regression. Our end goal was to provide a prediction regarding the likelihood of having type-2 diabetes. In addition to that, an implementation of a two-class decision forest model that had been constructed through the utilization of ensemble learning was carried out. The accuracy values of the decision forest, which they ultimately decided as their prediction model for type-2 DM, were pretty excellent. This was one of the reasons why they chose to use it. High values of prediction were achieved with F1-score values of 82.9%.

Following the evaluation of a number of different combinations of AdaBoost and XgBoost in addition to various combinations of ML techniques, Hasan et al.'s [15] the best accuracy rate of 95%. In this article, he presents a method for the prediction of diabetes that takes use of multi-layer perceptron's (MLP), a variety of classifiers, cross-validation, feature selection, missing values, outliers, and data normalization.

Dash and Vizhi [16] used ML to generate diabetic prediction among patients as potential answers to the difficulties that they were experiencing. In addition, a boosting methodology is made accessible for use with these two methodologies as part of the method.

An advanced deep learning model was published by Li et al. [17]. It is able to forecast glucose levels to a significant degree, which is the most significant factor that leads to the development of diabetes. Their model demonstrates competency by producing an ideal prediction with a low time lag in a dataset consisting of simulated patients as well as a dataset consisting of actual patients. This is the case in both datasets. In addition, the root-mean-square error (RMSE) method emerged victorious over the competition when compared to the alternatives that were investigated.

In a study that Amani Yahyaoui and colleagues carried out, they investigated the degree to which ML and DL algorithms differ in the accuracy of their diabetes prediction results [18]. A total accuracy of 83.67% RF, which consistently performed better than other testing methods when it came to diabetes categorization, was able to attain success in diabetic prediction. SVM achieved an accuracy of 65.38 percent in its predictions, whereas DL had a prediction accuracy of 76.81 percent.

Wang et al. [19] The relevance of unbalanced data with missing values is discussed in the predictive analysis of DM. They employed NB to normalize the data and compensate for missing values in their experiment. Over-sampling with the ADASYN algorithm is used to solve the problem of class imbalance. Finally, prediction is done using the RF. An experiment is carried out with the PIDD, and the outcomes are evaluated utilizing the combined technique of these classifiers, which works to enhance the findings individually.

Sajida Perveen and colleagues recommended the AdaBoost approach [20]. An AdaBoost ensemble technique performs better than bagging and J48 model for classifying diabetes patients. The author is driven to write about how crucial it is for healthcare professionals to predict and prevent diabetes mellitus because of the rising impact of diabetes on society globally. The author used model for patients of three distinct ages for the classification of diabetes in the Canadian population. To assess the effectiveness and accuracy, test data were used to three ensemble models (bagging, AdaBoost, and J48). In terms of accuracy, AdaBoost performs considerably better than its competitors, according to the findings. AdaBoost has the potential to enhance the accuracy of forecasts for a variety of other conditions, including coronary heart disease and hypertension, according to the scientists.

A method was suggested by Aliberti et al. [21], in which we evaluate the prediction models using glucose signal data from a sizable and diverse cohort of patients before applying them to predict future glucose levels in a new patient. When three literature techniques (FNN, RNN, and AR) are compared with two distinct types of solutions (Non-Auto Regressive NN and LSTM networks), the LSTM network outperforms all other models both for short-term and long-term forecasts.

Sierra-Sosa [22] developed a method to evaluate diabetes using LDA, SVM which provides accuracy of 92%, RNN with

94.6% of the evaluation of diabetes on variation parameters. A machine learning-based prediction algorithm created by Sneha and Gangil [23] searches for the best classifier to produce the next-best result. The method of choosing attributes and then performing early diabetes mellitus detection is known as "predictive analysis." The calculated value reveals that the DT's algorithm as well as RF's have the highest specificities for prediction, respectively of 98.20% and 98%.

Fikirte Girma et al. [24] used the R programming language and the Neural Network Back Propagation Model (J48, NB, and SVM based methods) to evaluate the accuracy of diabetes prediction. The accuracy of the Neural Network Back Propagation, J48, NB, and SVM algorithms, respectively, was 83.11%, 78.26%, 78.97%, and 81.69%.

A prediction system was developed by Ashikuzzaman and others [25] that uses the drop-down technique of data overfitting in the predictive model. In the work that was proposed, an advanced form of deep neural network was utilised, and it achieved an accuracy of 88.41%. Breast cancer and diabetes datasets were retrieved from the ML repository at UCI and used by Deepika Verma and her colleagues [26] to classify data using the NB, SMO, REP tree, J48, and MLP algorithms. After examining the effectiveness of every algorithm, J48 achieves a level of accuracy of 74.28% on the breast cancer dataset, while SMO achieves a level of accuracy of 76.80% on the diabetes data. Mohebbi and others [27] built multi-layer neural networks and CNN using the logistic regression model as a foundation. The author's dataset includes nine characteristics that are gathered for each patient. The total number of simulated days was 97,200 because each patient had data for 10,800 days. The attributes used in the analysis weren't properly discussed.

Ashok Kumar and others [28] developed a system for diagnosing diabetes by employing classification trees, SVM, LR, Na¨ıve Bayes, and other computational intelligence methods based on artificial neural networks. The results of the experiments showed that the accuracy of classification achieved by ANN and logistic regression was. 77% and 78%, respectively. Zeng et al. [29] For the analysis of uneven medical data, the authors recommend using the (SMOTE), which is a powerful data sampling algorithm. This will help ensure accurate results. In this specific inquiry, SMOOTH was compared to a total of eight different classifiers. These classifiers included SVM, BN, AdaBoost, K, C4.5, RBF Network, and LMT.

Lee and Kim [30] doing the deed of conormal subjects with and without type 2 diabetes were compared using 17 hyper-tri-glyceridemia waist (HW) measurements and individual anthropometric measurements in order to search for statistically significant differences between the two groups of people. NB and LR are two different machine learning models that are utilized to evaluate the predictive capacity of several phenotypes on the CPCSSN dataset, which has 667907 entries. This makes it possible to acquire findings that are more accurate and realistic based on the predictions that were created using the dataset. The results of experimental comparisons show that the AdaBoost ensemble approach outperforms both bagging and standalone J48 decision trees in terms of overall performance. This is proven by the conclusions of the experiments.

Veena Vijayan V and others, [31] concentrated on DM prediction with low error rates. Sets of data are retrieved from the UCI ML repository. Using the WEKA tool and MATLAB to verify accuracy, Diabetes can be predicted using AdaBoost-decision stump classifiers with an accuracy of 87.29%.

Saravana Kumar and others [32] built a system with Hadoop and the Map-Reduce methodology as the foundation. This model can predict the elements that put a person at risk for developing diabetes. Hadoop is the foundation of this system, which offers a cost-effective solution for any healthcare institution. Numerous clinics, laboratories, electronic health reports, and patient health reports all contributed significant amounts of data, which was collected. Hadoop was utilized to perform the processing, and the results were then dispersed among a number of servers in a manner that was determined by their geographical positions.

Lee and others [33] predicted fasting plasma glucose using Na¨ıve-bayes classifiers and logistic regression. In the LR and NB classifiers, female models had higher AUC than male models (females 74.1%-73.9% vs. males 68.6%-68.7%).

There is a model conceptualised by George and others [34] for the purpose of this research, six different scenarios containing the afore- mentioned variables have been selected, and the SVR model's tenfold cross-validation is being utilized in order to validate the suggested work. The selection of logical features and a suitable classifier is the most difficult task in machine learning : This experimental analysis aims to determine diabetes. Various classification techniques based on ML such as NB, SVM, LR, RF, K-NN and DT are employed as well as assessed on PIDD to forecast diabetes in a patient. Various measuring methods are used to assess the efficiency of all categorization systems.

III. METHODOLOGY USED

For categorizing data into distinct classes, classification algorithms are commonly employed in predictive analysis and maybe even pattern recognition. ANN as well as ML are advantageous technologies that can accomplish because of the power of their numerous categorization approaches. These technologies are widely employed in healthcare industry, where forecasting diagnosis is a difficult work due to increased missing values as well as discrepancies with in data collection [19]. Humans constantly learn from their previous experiences, while machines always obey human instructions. To train as well as create model in a specific field, you'll need to collect a large quantity of data, create a set of algorithms and test the model's correctness using multiple statistical metrics across properly and erroneously categorized examples [16], [28] . In this work, we going to use ML to create a model for diabetes prediction analysis that uses critical variables that are closely related to the condition. Figure 1 depicts the method used for creating and assessing the prediction models. Deepnote, a tool, was used to code in the Python programming language. With key integrated capabilities for application development, data visualization, and iterative data analysis, Deepnote provides a proven analytical and scientific Python package distribution.

The techniques for creating a model include various important phases that are discussed one by one, allowing the reasoning behind this study to be explored.

A. Dataset

We use Kaggle's PIDD from the National Institute of Diabetes, Digestive and Kidney Diseases (UCI) ML Repository was utilized for the experimental study, which offers useful attributes which are directly connected to this condition [35]. There are 768 entries in this dataset, 268 of which are expected to be diabetes patients, and 500 of which are projected to be non-diabetic patients, respectively, accounting for 34.9 percent and 65.1 percent of the total dataset. It has eight key traits and one result class, as shown in Figure 2 and this may be downloaded in CSV format.

B. Data pre-processing

As data is an integral part of model training, machine learning algorithms are fully reliant on it. When data is gathered in such haphazard manner from plenty of references, however there it seems to be a possibility of various divergences that the model will not be capable to tackle. As an outcome, pre- processing is necessary to remove any discrepancies as well as deliver a cleaned collected data. Replacing null values, computing innovative characteristics, partitioning data into training-testing phase, data encoding, normalizing data, and so on are all examples of this. Data unbalance, which occurs when one class includes so much occurrences than the other, seems to be another challenge that develops at pre-processing step [36].

1) Missing Values:

The value for various characteristics in the supplied sample is zero, resulting in missing data. To better appreciate this, consider the property diastolic blood pressure, which cannot have a zero value for a person [19]. The problem of missing values can be solved in two ways:

- record deletion.
- Imputation technique.

When the dataset is huge, the first approach, known as data deletion, is used. In this situation, you may eliminate records that have missing values, leaving a sufficient quantity of data for prediction. However, we're dealt with medical information, as well as the database utilized in this study has 768 records, which is a small number, and all of the characteristics are tightly connected. As a result, deleting entries with missing data is not a smart strategy in this circumstance. The imputation technique is the second strategy, which uses the most likely feature's class mean or group median to resolve missing data. Null value issues might be solved using the average of closest neighbors as well as the random number technique [36]. Missing values are handled in this study using the class mean feature.

2) Balance and unbalance dataset:

Data imbalance is an issue that happens frequently in the classification field, and it's the issue of disparities between favorable as well as unfavorable projected classes If the number of favourable and adverse observations in a dataset are equal, then the dataset is said to be balanced; otherwise, it is said to be imbalanced. Evaluating a balanced dataset is straightforward because there is no space for bias. If the proportion of positive samples in the provided dataset is 10%, then the classifier's accuracy in predicting the entire negative is 9%. Even though the accuracy is unquestionably quite excellent, it can occasionally be misleading. Random sampling and oversampling are two techniques that can be employed when faced with an uneven data distribution. Other performance metrics, such as specificity, sensitivity (recall), F-score, and precision, are employed to resolve the issue of unbalanced data. Over-sampling is the process of recreating the underclass without sacrificing any of the information contained in the initial training dataset. It has the potential to be overfit despite this. In the technique of under-sampling, the upper class is eliminated so that the dataset is more evenly distributed; however, this may result in the loss of significant data [19]. In order to address the issue of class imbalance discovered during this research, the oversampling strategy was implemented.

3) Data normalization:

During the pre-processing step, it is quite important. If you have a dataset, you might be able to include characteristics of various measures [36]. Because certain characteristics in the supplied dataset are on a short- range scale and others are on a large-range scale, normalization is the process of putting all on the same scales as well as measures for simple comparison. The most often used data normalization approaches are z-score as well as min-max. The min-max approach is utilized in this study.

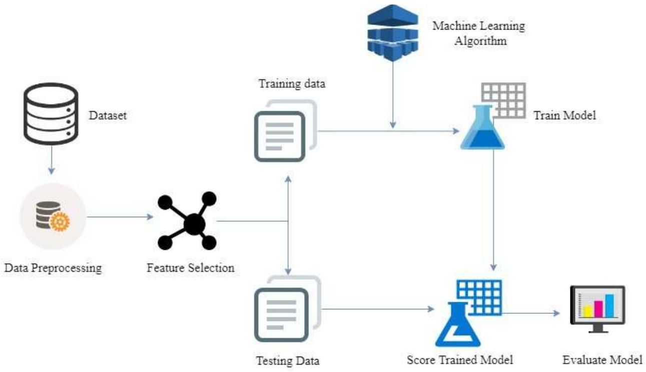
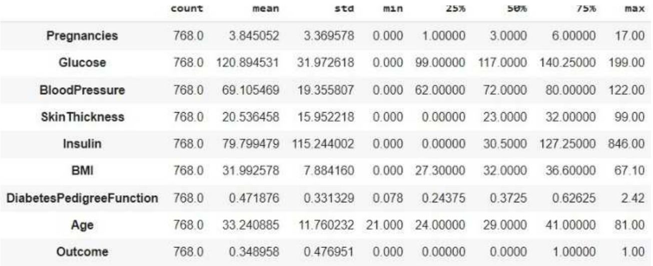


Fig. 1. Process followed for evaluation of algorithms



	count	mean	std	min	25%	50%	75%	max
Pregnancies	768.0	3.845052	3.369578	0.000	1.00000	3.0000	6.00000	17.00
Glucose	768.0	120.894531	31.972618	0.000	99.00000	117.0000	140.25000	199.00
BloodPressure	768.0	69.105469	19.355807	0.000	62.00000	72.0000	80.00000	122.00
SkinThickness	768.0	20.536458	15.952218	0.000	0.00000	23.0000	32.00000	99.00
Insulin	768.0	79.799479	115.244002	0.000	0.00000	30.5000	127.25000	846.00
BMI	768.0	31.992578	7.884160	0.000	27.30000	32.0000	36.60000	67.10
DiabetesPedigreeFunction	768.0	0.471876	0.331329	0.078	0.24375	0.3725	0.62625	2.42
Age	768.0	33.240885	11.760232	21.000	24.00000	29.0000	41.00000	81.00
Outcome	768.0	0.348958	0.476951	0.000	0.00000	0.0000	1.00000	1.00

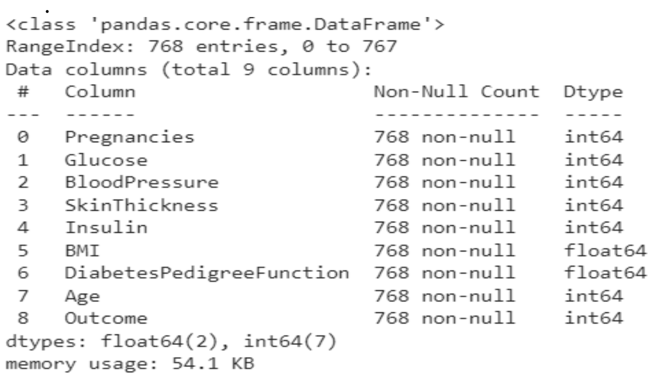
Fig. 2. Attributes and their characterstics of PIDD

C. Techniques used in Forecasting Analysis

This study employ's six categorization methods, each of which is discussed separately.

1) Logistic Regression (LR):

The goal of logistic regression is to determine the optimal line between the two groups. It is commonly used in issues involving linear classification. It is comparable to Linear Regression in terms of linearity [37]



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype  
---  ------                    --------------  -----  
 0   Pregnancies               768 non-null    int64  
 1   Glucose                   768 non-null    int64  
 2   BloodPressure             768 non-null    int64  
 3   SkinThickness             768 non-null    int64  
 4   Insulin                   768 non-null    int64  
 5   BMI                       768 non-null    float64
 6   DiabetesPedigreeFunction  768 non-null    float64
 7   Age                       768 non-null    int64  
 8   Outcome                   768 non-null    int64  
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
```

Fig. 3. Description of Diabetes Dataset

2) K-Nearest Neighbour (KNN):

KNN predicts the value to be anticipated based on the class density of the vector created by the independent variables. It is estimated the distance between the projected point and other points. This is done with the Minkowski distance function. (K: We shall calculate the number of closest neighbors.) [37]

The working steps are listed below in KNN [38]:

- Both dataset as well as classes of the training set are uploaded during the model's training stage.
- The value of K is chosen in the second stage. The k variable indicates how many neighbors are involved in the majority voting procedure. The K value is utilized to transform an unidentified class into a specified class based on Euclidean distance or elbow function measurements.

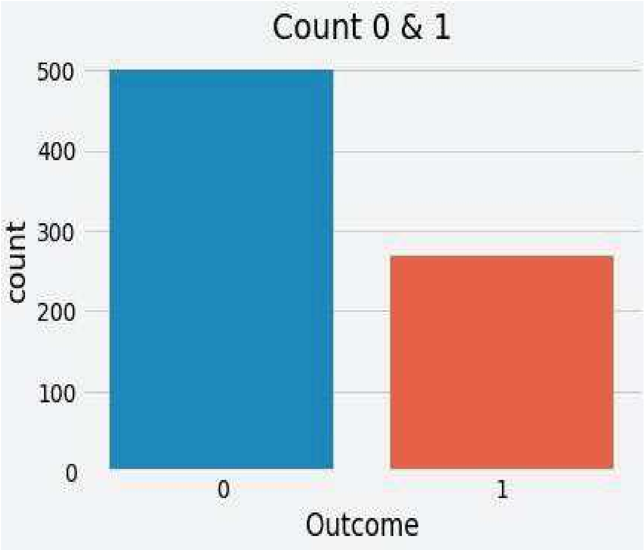


Fig. 4. Outcome of diabetic and non-diabetic patients

3) Support Vector Machine (SVM):

For the purposes of area regression and classification, SVM is a supervised computer model that is utilised extensively. Using support vector machines (SVM), the data item is plotted in a space with a greater dimension. Assume it represents data in n-dimension space, if you have 'n' features. The support vector machine (SVM) is used to build the hyperplane connecting datasets that most effectively segregates the dataset into classes. The difficult challenge is to choose the best hyperplane in dimensional space, with the appropriate hyperplane being the plane with the greatest difference between two classes. The support vectors are the locations that are nearest to the hyper- plane. The items are mapped according to the hyperplane's set limits. The new sample's class usually determined by a hyperplane which corresponds to one of the classes running parallel to it. [38].

4) Na¨ıve Bayes (NB):

The NB algorithm is a supervised learning model relying on Bayes' Theorem as well as the predictor independence assumption. An NB model states that the participation of subsequent characteristics in the same class is unaffected by the existence of one character in the class [37].

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \qquad (1)$$

P(A|B): The chance that event A will happen if event B happens.

- Create 'n' training sets by randomly selecting data from the original data sets and replacing them.
- For the 'n' training sets, they are needed to construct 'n' decision tree models. Without trimming, the DT will expand to its full potential.
- The optimal feature for each unique decision tree is chosen in accordance with the Gini value for each division.
- For the node's training samples to all belong to the same class, repeat step 3 as necessary.
- The final classification outcomes for the classification issue are decided by the voting of all 'n' DT's, which together make up a random forest.

5) Decision Tree (DT):

A DT is a supervised learning approach that is commonly used in resolving categorization issues. It begins as a single node and grows into a tree. It constructs a model for predicting the value of the parameter by extraction and learning basic rules from data attributes [37].

- To begin, load the data for training, which contains 'm' characteristics and demonstrates the dataset's performance.
- Bagging is a technique for randomly sampling a fraction of training (with alternative) and selecting 'n' characteristics from 'm' characteristics.
- The 'n' training characteristics are employed in the DT analysis.
- In each DT, the Gini index is utilized to pick the splitting nodes.
- The previous procedures will be repeated for predicting an arbitrary number of DT's.
- In forecasting the target class, the majority voted class is computed from the total number of votes cast by all trees.
- In the case of classification, use the mode of all predictions, and in regression's case, we use mean.

6) Random Forest (RF):

RF is a powerful supervised learning method that may be applied to a variety of different types of problems, including those involving classification and regression. A large number of DTs are included in the ensemble classifier that was used to make this prediction [38], and the majority of the votes that were cast came from those trees. It generates better results than individual decision tree classifiers do, which is a significant advantage. It generates better results than individual decision tree classifiers. It trains each tree using the bagging approach, which generates random of tests are carried out to acquire statistically valid results. If K is set to ten, the test will be repeated ten times. Out of K iterations, one part is chosen as testing set as well as the rest K-1 sections are chosen as a training set for each value of K. The advantage of this technique is that each section has the same possibility of becoming a test set. Consider the average of the data acquired after K trials, which represent the model's performance metrics [19], [38]. Equation 1 gives the approximate average defect of the k tests. sample features in the provided data. ID3 and CART are two extensively used methods for creating decision trees. The suggested technique employs the CART algorithm, and the Gini index serves as the appropriate segmentation index. The following are some useful RF steps [19]:

D. Approach for evaluation

The model's efficiency and reliability are assessed using -fold cross-validation. To verify the efficiency, the existing dataset is converted to a train-test set denotes the number of parts that make up the entire data item. Several repetitions of tests are carried out to acquire statistically valid results. IfK is set to ten, the test will be repeated ten times. Out ofK iterations, one part is chosen as testing set as well as the rest K-1 sections are chosen as a training set for each value of K. The advantage of this technique is that each section has the same possibility of becoming a test set. Consider the average of the data acquired after K trials, which represent the model's performance metrics [19], [38]. Equation 1 gives the approximate average defect of the k tests.

$$E = \frac{1}{k}\sum_{i=1}^{k} E_i \qquad (2)$$

Where Ei is the error achieved for each of the test dataset's passes.

TABLE I. CLASSIFICATION PERFORMANCE OF ALGORITHMS BASED ON DIFFERENT MEASURES.

Algorithms	TN	FP	FN	TP	Precision	Recall	F1-score	Accuracy
LR	79	20	10	45	69.23	81.82	75	80.52
KNN	84	15	12	43	74.14	78.18	76.11	82.47
SVM	83	16	11	44	73.33	80	76.52	82.47
NB	73	26	11	44	62	80	70.4	75.97
RF	88	11	8	47	81.03	85.45	83.19	87.66
DT	82	17	13	42	71.19	76.36	73.68	80.52

Effectiveness of the various classifiers used to create a model is measured using certain major statistical indicators. Sensitivity (recall), accuracy, precision, F-score, and specificity are the measurements. Classification labels are utilized to compute all parameters [38].

1) *Accuracy:* It is the proportion of correctly categorized values.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \qquad (3)$$

2) *Sensitivity / Recall (R):* It demonstrates "How many TPs have been appropriately defined"?

$$Sensitivity = \frac{TP}{TP+FN} \qquad (4)$$

3) *Specificity:* It is calculated using the method below to determine the genuine negative rate.

$$Specificity = \frac{TN}{FP+TN} \qquad (5)$$

4) *Precision (P):* It demonstrates "how many of the positivenumbers we predict are TP" ?

$$Precision = \frac{TP}{TP+FP} \qquad (6)$$

5) *F1-score*: It's the harmonic mean of Precision as well as Sensitivity. The F-score becomes essential if both Precision and Recall scores are crucial for the task. Sensitivity (recall), accuracy, precision, F-score, and specificity are the measurements. TP (True Positive), TN (True Negative), FP (False Positive), and FN (False Negative) classification labels are utilized to compute all parameters [38].

$$F1 - Score = \frac{2*P*R}{P+R} \qquad (7)$$

IV. EXPERIMENTAL RESULTS AND ANALYSIS

In this work, a model that assists in the earlier identification of diabetes by focusing on essential disease characteristics is created utilising six different classifier algorithms. These techniques are employed to develop a model. The methods of LR, KNN, SVM, NB,RF, and DT were applied, and the dataset used was PIDD, which was obtained from the UCI ML Repository [35]. Create a system for handling the data during the trials and organise it. In order to tackle the issue of class unbalance, the null values are replaced with the average of the features class, and the over-sampling method is used as a solution to the problem. The formula for determining how accurate a measurement is can be found in Equation 3. Accuracy, sensitivity (recall), specificity, and F-score are some of the major performance criteria that have been evaluated. Figure 3 presents a summary of the dataset for your perusal. It consists of 768 instances and eight features with a single class label, where a value of '0' indicates patients who do not have diabetes and a value of '1' indicates patients who do have diabetes. The entire thing is organised into a single table. Figure 4 depicts the overall outcomes of both diabetic and non-diabetic patients, with a total of 268 diabetic patients and 500 non-diabetic patients included. There are both tabular and graphical representations of the results that can be seen here. Table I provides a summary of the findings from all of the classifiers and includes all of the performance indicators that are necessary to evaluate the efficacy of each classifier. The individual performance of an algorithm is represented here by Figure 5, which is based on statistical indicators. According to the data, the Random Forest (RF) classifier has the highest accuracy, coming in at 87.66%. As a result, it is adequate for making predictions regarding diabetes mellitus using our model.

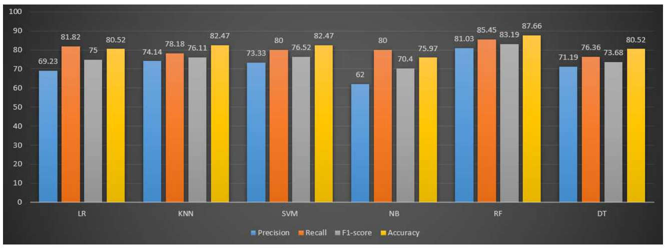


Fig. 5. Comparative Study of Different ML Models based on Statistical Indicators.

V. CONCLUSION

Diabetes Mellitus is a real-world nasty illness, as well as prior detection is always a challenge. This Work uses ML classification methods to create a model that effectively addresses all of the problems and is useful in the early detection of diabetic illness. The tests are carried out on the PIDD dataset utilizing six ML Algo's: K-NN, LR, NB, SVM, RF, and DT. The collection of data contains 768 records as well as eight key diabetes-related variables, as well as a class label that indicates the result of diabetic and non-diabetic individuals. Our major goal is to improve the model's accuracy, but we've also looked at other critical performance measures including recall, precision, F-score, and specificity. Confusion measures such as TP, TN, FP, and FN are utilized to analyze these performance parameters. According to the data, Random Forest (RF) outperforms the other employed classifiers and obtains the highest level of accuracy, approximately 88%. In light of this, the NB classifier is an outstanding choice for our model.

The magnitude of the dataset and the absence of attribute value information are two of the limitations of this study. For our diabetes prediction model to be 99.99% accurate, we will need thousands of data with no incomplete information. Future work will primarily concentrate on incorporating additional methods, such as Deep Learning (DL) Techniques, into the current model in order to improve the model's predictive accuracy. The evaluation of the models on a massive dataset with very few or no missing attribute values will lead to the discovery of new things and an enhancement in the predictive ability of the models.

REFERENCES

[1] Global report on diabetes, 2016. Available at: https://apps.who.int/iris/bitstream/handle/10665/204871/9789241565257eng.pdf ; jsessionid = 2BC28035503CFAFF 295E70CFB4A0E1DF ?Sequence = 1.

[2] Diabetes: Asia's 'silent killer, november 14, 2013.Available at: Di Cataldo, Edoardo Patti, and Andrea Acquaviva. A multi-patient data-driven approach to blood glucose prediction. IEEE Access, 7:69311–69325, 2019.www.bbc.com/news/world-asia-24740288.

[3] Emerging Risk Factors Collaboration et al. Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies. The Lancet, 375(9733):2215– 2222, 2010.

[4] How is the pancreas linked with diabetes? by Jenna Fletcher. Available at: https://www.medicalnewstoday.com/articles/325018how-is-the- pancreas-linked-with-diabetes. Published: April 23, 2019. Accessed May 10,2022.

[5] Suyash Srivastava, Lokesh Sharma, Vijeta Sharma, Ajai Kumar, and Hemant Darbari. Prediction of diabetes using artificial neural network approach. In Engineering Vibration, Communication and Information Processing, pages 679–687. Springer, 2019.

[6] Bayu Adhi Tama and Kyung-Hyune Rhee. Tree-based classifier ensembles for early detection method of diabetes: an exploratory study. Artificial Intelligence Review, 51(3):355–370, 2019.

[7] Gopi Battineni, Getu Gamo Sagaro, Nalini Chinatalapudi, and Francesco Amenta. Applications of machine learning predictive models in the chronic disease diagnosis. Journal of personalized medicine, 10(2):21, 2020.

[8] Ioannis Kavakiotis, Olga Tsave, Athanasios Salifoglou, Nicos Maglaveras, Ioannis Vlahavas, and Ioanna Chouvarda. Machine learning and data mining methods in diabetes research. Computational and structural biotechnology journal, 15:104–116, 2017.

[9] Samrudhi R Kaware and Vinod S Wadne. Improve the performance of cancer and diabetes detection using novel technique of machine learning. Technical report, EasyChair, 2020.

[10] Usama Ahmed, Ghassan F. Issa, Muhammad Adnan Khan, Shabib Aftab, Muhammad Farhan Khan, Raed A. T. Said, Taher M. Ghazal, and Munir Ahmad. Intelligent decision support model using tongue image features for healthcare monitoring of diabetes diagnosis and classification. IEEE Access, 10:8529–8538, 2021.

[11] S.N. Deepa and Abhishek Banerjee. Prediction of diabetes empowered with fused machine learning. Network Modeling Analysis in Health Informatics and Bioinformatics, 10:1–16, 2022.

[12] Nikos Fazakis, Otilia Kocsis, Elias Dritsas, Sotiris Alexiou, Nikos Fakotakis, and Konstantinos Moustakas. Machine learning tools for long-term type 2 diabetes risk prediction. IEEE Access, 9:103737–103757, 2021.

[13] S Sivaranjani, S Ananya, J Aravinth, and R Karthika. Diabetes prediction using machine learning algorithms with feature selection and dimensionality reduction. In 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), volume 1, pages 141–146. IEEE, 2021.

[14] Asif Hassan Syed and Tabrej Khan. Machine learning-based application for predicting risk of type 2 diabetes mellitus (t2dm) in saudi arabia: A retrospective cross-sectional study. IEEE Access, 8:199539–199561, 2020.

[15] Md. Kamrul Hasan, Md. Ashraful Alam, Dola Das, Eklas Hossain, and Mahmudul Hasan. Diabetes prediction using ensembling of different machine learning classifiers. IEEE Access, 8:76516–76531, 2020.

[16] Dr.Kayal Vizhi and Aman Dash. Diabetes prediction using machine learning. International Journal of Advanced Science and Technology, 29(06):2842 – 2852, May 2020.

[17] Kezhi Li, John Daniels, Chengyuan Liu, Pau Herrero, and Pantelis Geor- giou. Convolutional recurrent neural networks for glucose prediction. IEEE Journal of Biomedical and Health Informatics, 24(2):603–613, 2020.

[18] Amani Yahyaoui, Akhtar Jamil, Jawad Rasheed, and Mirsat Yesiltepe. A decision support system for diabetes prediction using machine learning and deep learning techniques. In 2019 1st International Informatics and Software Engineering Conference (UBMYK), pages 1–4, 2019.

[19] Qian Wang, Weijia Cao, Jiawei Guo, Jiadong Ren, Yongqiang Cheng, and Darryl N Davis. Dmp mi: an effective diabetes mellitus classification algorithm on imbalanced data with missing values. IEEE Access, 7:102232–102238, 2019.

[20] Sajida Perveen, Muhammad Shahbaz, Karim Keshavjee, and Aziz Guergachi. Metabolic syndrome and development of diabetes mellitus: Predictive modeling based on machine learning techniques. IEEE Access, 7:1365–1375, 2019.

[21] Alessandro Aliberti, Irene Pupillo, Stefano Terna, Enrico Macii, Santa Di Cataldo, Edoardo Patti, and Andrea Acquaviva. A multi-patient data-driven approach to blood glucose prediction. *IEEE Access*, 7:69311–69325, 2019

[22] Daniel Sierra-Sosa, Begonya Garcia-Zapirain, Cristian Castillo, Ibon Oleagordia, Roberto Nun˜o-Solinis, Maider Urtaran-Laresgoiti, and Adel Elmaghraby. Scalable healthcare assessment for diabetic patients using deep learning on multiple gpus. IEEE Transactions on Industrial Informatics, 15(10):5682–5689, 2019.

[23] N. Sneha and Tarun Gangil. Analysis of diabetes mellitus for early prediction using optimal features selection. Journal of Big Data, 6(13), 2019.

[24] Fikirte Girma Woldemichael and Sumitra Menaria. Prediction of diabetes using data mining techniques. In 2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI), pages 414–418, 2018.

[25] Akm Ashiquzzaman, Abdul Kawsar Tushar, Md. Rashedul Islam, Dongkoo Shon, Kichang Im, Jeong-Ho Park, Dong-Sun Lim, and Jongmyon Kim. Reduction of overfitting in diabetes prediction using deep learning neural network. In Kuinam J. Kim, Hyuncheol Kim, and Nakhoon Baek, editors, IT Convergence and Security 2017, pages 35–43, Singapore, 2018. Springer Singapore.

[26] Deepika Verma and Nidhi Mishra. Analysis and prediction of breast cancer and diabetes disease datasets using data mining classificationtechniques. In 2017 International Conference on Intelligent Sustainable Systems (ICISS), pages 533–538, 2017.

[27] Ali Mohebbi, Tinna B. Arado´ttir, Alexander R. Johansen, Henrik Bengts- son, Marco Fraccaro, and Morten Mørup. A deep learning approach to adherence detection for type 2 diabetics. In 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pages 2896–2899, 2017.

[28] Dwivedi and Ashok Kumar. Analysis of computational intelligence techniques for diabetes mellitus prediction. Neural Computing and Applications, 30(12), 2018.

[29] Min Zeng, Beiji Zou, Faran Wei, Xiyao Liu, and Lei Wang. Effective prediction of three common diseases by combining smote with tomek links technique for imbalanced medical data. In 2016 IEEE International Conference of Online Analysis and Computing Science (ICOACS), pages 225–228, 2016.

[30] Bum Ju Lee and Jong Yeol Kim. Identification of type 2 diabetes risk factors using phenotypes consisting of anthropometry and triglycerides based on machine learning. IEEE Journal of Biomedical and Health Informatics, 20(1):39–46, 2016.

[31] V. Veena Vijayan and C. Anjali. Prediction and diagnosis of diabetes mellitus — a machine learning approach. In 2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS), pages 122–127, 2015.

[32] N.M. Saravana kumar, T. Eswari, P. Sampath, and S. Lavanya. Predictive methodology for diabetic data analysis in big data. Procedia Computer Science, 50:203–208, 2015. Big Data, Cloud and Computing Challenges.

[33] Bum Ju Lee, Boncho Ku, Jiho Nam, Duong Duc Pham, and Jong Yeol Kim. Prediction of fasting plasma glucose status using anthropometric measures for diagnosing type 2 diabetes. IEEE Journal of Biomedical and Health Informatics, 18(2):555–561, 2014.

[34] Eleni I. Georga, Vasilios C. Protopappas, Diego Ardigo`, Michela Marina, Ivana Zavaroni, Demosthenes Polyzos, and Dimitrios I. Fotiadis. Multivariate prediction of subcutaneous glucose concentration in type 1 diabetes patients based on support vector regression. IEEE Journal of Biomedical and Health Informatics, 17(1):71–81, 2013.

[35] Uci machine learning repository. https://www.kaggle.com/datasets/uciml/pima- indians-diabetes-database.

[36] Brian Malley, Daniele Ramazzotti, and Joy Tzung-yu Wu. Data preprocessing. Secondary analysis of electronic health records, pages 115–141, 2016.

[37] Esma Bozkurt. Machine learning classification algorithms with codes. https://medium.com/analytics-vidhya/machine-learning-classificationalgorithms-with-codes-5a8af4491fcb, 2021.

[38] Gaurav Tripathi and Rakesh Kumar. Early prediction of diabetes mellitususing machine learning. In *2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO)*, pages 1009–1014. IEEE, 2020.