Prediction of Diabetic condition using Different Machine Learning Approaches and Different Datasets

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/ICCPCT58313.2023.10245246

Abstract*— Diabetes is a insulin hormone disorder in our bodies, which, when it doesn't work well or efficiently, causes our blood sugar levels to rise. Numerous people experience high blood sugar levels, which have a detrimental impact on other human organs. An unhealthy lifestyle, poor eating habits, and a lack of exercise are the root causes of this condition. Other complications like coronary failure, blindness, urinary organ disease, etc. will also result from this. Diagnostic facility are used to diagnose the disease and for health reports. This results in a time and money loss in the process. This paper explains that machine learning techniques could be used instead to detect diabetes quickly with accuracy so early detection would help control its aftereffects through preventive measures taken at the right time while saving both money and time compared to traditional methods.

Keywords— Diabetes Disease Prediction, Support Vector Machine, Machine Learning, Random Forest, Decision Tree

I. Introduction

Diabetes is classified as the most hazardous degenerative illnesses since it causes other significant complications [1, 2]. Diabetes ailments are also known as hypoglycemia, a term that refers to several metabolic abnormalities [3]. Diabetes could lead to many different complications, including cardiovascular ailments, astigmatism, renal failure, etc [4]. Another chronic condition that is one of the worst illnesses is diabetes, which generally refers to a major metabolic problem. There are two different kinds of diabetes: category 1 Mellitus and category 2 Mellitus. When there is an insufficient amount of glucose produced in the body, a person is said to have type 1 diabetes. In the second category, the body is either unable to make optimal use of the glucose that it makes or it does not release a enough amount of glucose into the circulatory system. Ailments that fall under group 1 are more common in children and teenagers, but adults can also be affected by them. In general, glucose is safer for people who do not have category 2 when compared to others who do not have it. As a potential treatment for type 1 diabetes, glucose might be injected directly into the patient's muscle tissue, just below the epidermis. Category 2 Diabetes, on either hand, could be healed by following a healthful diet, counting calories, as well as daily exercise. Several ailments could be avoided unless Diabetes was detected in its initial phases. Because of current technology advancements in IoT, Artificial Intelligence (AI), and Blockchain in the present medical system, early intervention, and forecasting of ailment is now achievable. AI has ushered in a fundamental change in Diabetes management, moving away from conventional planning processes and toward data-driven accuracy treatment. The Internet of Things (IoT) provides a connected environment for the intelligent healthcare system. Reinforcement learning and machine learning were AI-based approaches. In the pharmaceutical and healthcare industry, machine learning improves efficiencies and reduced medical costs.

For the purpose of diagnosing diabetes and determining its prognosis, a variety of publications based on data extraction or machine learning was available. The utilization of data mining and machine learning strategies would be crucial to the accomplishment of their objectives. Methods of data collection are required in order to extract rules and patterns from a large amount of information regarding diabetes; yet, computer learning and the discovery of patterns were certainly vital for the process of learning and automating the system. Several different machine learning strategies were utilized in the Mellitus treatment in order to generate trend indicators.. neural network, Decision Tree (DT) classification and K-Mean, regression trees, and Principal Component Analysis (PCA) based algorithms for better diabetes treatment are few examples of these [5]. There is an urgent need to diagnosing, forecasting, and managing blood sugar levels to decrease the fatality of this disease.

Diabetes of the Category 2 variety was a severe ailment. Diabetes is one of the most dangerous chronic disorders since it leads to serious problems as the disease progresses. Diabetes refers to a collection of metabolic disorders and is also the name given to the conditions that fall under this umbrella term. Diabetes could lead to a range of consequences, such as cardiovascular disorders, renal failure, blindness, and so on and so forth. Diabetes, which is typically characterized by high blood glucose levels, should be considered a persistent condition that ranks among the deadliest diseases. Diabetes can be broken down into two categories: category 1 and category 2 Mellitus. Diabetes type 1 occurs due to pancreas defficiency in producing sufficient amount of insulin. Diabetes type 2 is characterized by either an inability of the body to utilize the glucose it generates in an effective manner or an inadequate release of glucose into the circulatory system.

Although type 1 illness is most common in children and adolescents, adults are not immune to the risk of developing the condition. When compared to persons who do not have category 2 Diabetes, category 2 Diabetes was usually milder. Glucose could be injected into the patient's fat deposits beneath the epidermis to treat category 1 Diabetes. However, it could be healed by following a nutritious diet, losing weight, and daily exercise. Several ailments could be avoided if Diabetes was indeed detected in its initial phases. Because of current technology advancements in IoT, Artificial Intelligence (AI), and Blockchain in the present medical system, early detection and forecasting of ailments is now probable. AI has ushered in a fundamental change in Diabetes care, drifting away from traditional planning processes and toward data-driven accuracy treatment. The Internet of Things

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

(IoT) provides a connected environment for the intelligent medical system. Reinforcement learning and machine learning were AI-based approaches. In the healthcare sector, machine learning can improve efficiencies and reduce medical costs. Various literature based on statistical extraction and machine learning is accessible for glucose diagnosis and prognosis. The utilization of data mining and machine learning strategies would be crucial to the accomplishment of their objectives. Data gathering techniques are important for isolating rules and patterns from a massive amount of diabetic information, and computer learning and pattern recognition are important for both automating the system and learning new information. Learning from massive amounts of diabetic information is important. Several approaches to automated glucose diagnosis, prognosis, or treatment via ML and AI have been provided in only two categories of learning, supervised and unsupervised [6].

II. Literature Survey

H. Suriya Babu et al [17], in this research, they have created a model for analyzing the dataset by making use of MapReduce and Hadoop techniques. The algorithm that is being created will be able to forecast the kind of diabetes as well as the risk that is associated with it. Both the Naive Bayes algorithm and the decision tree algorithm have been utilized by them. The fact that it is based on Hadoop makes it a cost-effective solution for society. Both the effectiveness of both approaches and the performance of both algorithms were evaluated and compared. They discovered the patterns that were concealed inside the dataset by employing a decision tree. In the case of logistic regression, the accuracy is 96%. AdaBoost was the most accurate model, coming in at 98.8% accuracy overall.

In this particular research project, carried out by Ayman Mir and Sudhir N. Dhage [14], the authors used the WEKA tools to develop their model. For research that is driven by data, this machine learning and data mining toolset is quite well-known. WEKA 3.28 is the version that is being utilized here. They decided to make use of this toolkit due to the fact that it effectively analyzes the performance of a model and makes it possible to compare data in real time. Several other algorithms, including SVM,NB andRF were used. They have created a training/test dataset. Comparisons are made between the training times, testing times, and accuracy values of the various methods. With an accuracy value of 0.7913, the SVM demonstrated the highest level of precision.

Researchers Mitushi Soni and Dr. Sunita Varma [15] conducted this study, in which they attempted to forecast diabetes by experimenting with classification and an ensemble of algorithms. Their objective is to develop a model that can diagnose diabetes with a higher degree of precision. The SVM algorithm, k-nearest neighbor algorithm, RF algorithm, DT algorithm, logistic regression, and gradient boosting are the proposed methods that have been utilized in this study. They have eliminated every instance that had the number zero since it is mathematically impossible. The dataset undergoes data pre-processing since it is a critically crucial step that improves both the model's accuracy and its ability to make accurate predictions. They chose to utilize an ensemble of machine learning algorithms because, in comparison to using individual algorithms, the ensemble provides a better level of accuracy of 77%.

They evaluated diabetes risk factors using a dataset from epidemiological population study. In an article [13], researcher conducted the research. The range of ages that make up the average is from 24 to 64 years. Machine learning will be able to detect otherwise healthy persons who have a significant probability of developing type 2 diabetes. Recall, accuracy, and specificity are the metrics that they have utilized in order to evaluate the efficacy of their model. In order to build their model, they utilized the Support Vector Machine technique. They made the observation that adding additional features did not result in an increase in performance or an improvement in performance. High glucose level reported at the two-hour mark in the OGTT can be taken as a significant indicator of the probability of developing type-2 diabetes.

In this study by K. VijiyaKumar et al [7], they have gathered their dataset from the database. They have employed techniques such as data cleaning, integration, and transformation as part of the pre-processing steps. They have decided to use random forest as the basis for their model because, in comparison to other machine learning algorithms, it provides a greater level of accuracy. The process of identifying and eliminating erroneous or corrupted data from a dataset is referred to as "data cleaning." The process of converting numerical or alphabetical digital data into a more organized or simplified form of data is referred to as "data reduction." The research properly early predicts an individual's risk of developing diabetes. The random forest algorithm has a precision that is more than 90% of the time.

[18], the authors of this research, H. Suriya Babu et al. suggest a system that centers on making use of the AdaBoost algorithm for its model. They have experimented with a variety of basic classifiers for this approach. The implementation of their system takes place in four stages. During the training phase, they utilized a global dataset, and during the testing phase, they utilized a local dataset. They have used sensitivity, error rate, and specificity as the performance parameters that they have been looking at. They have validated the accuracy of their results by using the Weka interface. The Decision Stump had the highest accuracy out of the four different classifiers that were utilized, coming up at 80.72%. The accuracy of the decision tree comes in at 77.6%, the accuracy of the Support Vector Machine comes in at 76.987%, and the accuracy of the Nave Bayes method comes in at 79.687%. These results came from various algorithms. The Decision Stump method, with its 19.27% error rate, is also capable of achieving the lowest error rate. There is a mistake rate of 22.39 percent when using the DT method. The error rate for the SVM method is 20.31 percent, and the error rate for the Naive Bayes algorithm is also 20.31 percent.

Machine learning models were trained and tested using publically available datasets in this study by M. D. Kamrul Hasan and others. They obtained the datasets from the Pima datasets, which had a total of 768 data points and nine distinct attributes each. The model returns a value of either zero or one as its result. One suggests that the person does not have diabetes, while the other suggests that the person does have diabetes based on the results of the test. They utilized a variety of machine learning strategies, including the decision tree method, the Nave Bayes algorithm, Logical Regression, the SVM algorithm, the Random Forest (RF) methodology, the k-nearest neighbor algorithm, and the AdaBoost algorithm.

In this article by Dr. Kayal Vizhi and Aman Dash [9], the project model can only evaluate one specific or particular parameter. It does not take into account any of the other parameters. This piece of writing requires additional editing. They obtained the datasets from the Pima datasets, which had a total of 768 data points and nine distinct attributes each. The model returns a value of either zero or one as its result. One suggests that the person does not have diabetes, while the other suggests that the person does have diabetes based on the results of the test. Their concept is entirely reliant on a framework that is built on the Internet of Things as well as cloud computing. They have placed a strong priority on protecting the privacy and well-being of the individual. They have employed a feature selection process, as well as gradient boost and a logical regression approach. They came to the conclusion that logical regression is more accurate than other algorithms, achieving an accuracy of 78%, and they stated this as their conclusion.

Jobeda This study made use of a number of different algorithms, including those developed by Jamal Khanam and Simon Y. Foo [10], the decision tree method, the Naive Bayes, the Logical Regression, the SVM, the RF, the KNN algorithms, and the AdaBoost algorithm. They obtained the datasets from the Pima datasets, which had a total of 768 data points and nine distinct attributes each. The model returns a value of either zero or one as its result. One suggests that the person does not have diabetes of the test, while the other suggests that the person does have diabetes based on the results of the test. They were able to get rid of three of the nine parameters by employing pre-processing strategies. In order to ensure that the models are accurate, the Weka tool is utilized. The accuracy of every model was greater than 70%. The ANN algorithm achieved an accuracy of 88.57%, which was the highest of any method.

Anjali C and Veena Vijayan V [8], the purpose of their research, as outlined in this work, is to identify diabetes at an earlier stage by making use of a variety of machine learning approaches, and to enhance productivity. They have implemented both principal component analysis and k-means algorithms in the system that they have suggested. The dataset that was utilized is considered to be in the public domain. They have constructed a number of models by utilizing the k-means, ANN, and SVM techniques. While training the models, a dataset that has already been processed will be used. They built a confusion matrix and put the user's data through its paces to see how well it performed. New data can be annotated with K-cluster centroids if necessary. The centroid needs to be updated after each new evaluation that is performed. They have loaded the CSV data with the help of Panda data. They came to the conclusion that the k-means method provides a higher level of accuracy when compared to other algorithms..

III. Datasets used for experiments

Our analysis' main goal is to determine whether the person has diabetes. Private patient data from Bangladesh and PIMA diabetes data were used in this study. This study divided the sample into two categories: training (70%), and testing (30%). Using the PIDD medical dataset, this investigation was completed. There are 768 entries and 9 attributes in the database for schools in PIMA data (0 or 1) 500 participants in class (0) do not have diabetes, compared to 268 persons in class (1) [9]. Table 1 summarises the PIMA diabetes dataset.

Table 1. PIMA dataset attributes

Attribute Type	Attribute Description
All Numeric	Pregnancy frequency, Plasma glucose concentration in two hours duration, Blood pressure (mm Hg) Skin thickness (mm) Insulin (mu U/ml) BMI (weight in kg / (height in square m) Pedigree function Age (years),
Nominal	Class variable (diabetic or not)

BanglaDesh Dataset [11] was obtained in 2019 utilising direct questions from Sylhet Diabetes Hospital patients in Sylhet, Bangladesh, and was approved by a clinician.

Table 2. The Bangladesh diabetes dataset Attributes

Attribute Used	Attribute Type
Age, Gender, Polyuria Polydipsia sudden weight loss weakness Polyphagia Genital thrush visual blurring Itching Irritability delayed healing partial paresis muscle stiffness Alopecia Obesity	All binary except age which is Continuous
class	Binary Value

IV. Methodology

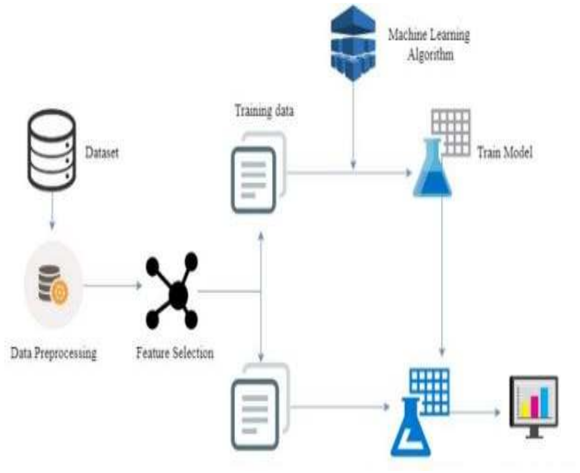

Fig 1 Working of the model

A. Data Pre-Processing

Preprocessing aids in determining which machine learning technique to generate will be more useful. The removal of outliers, the filling in of missing values, data duplication, and characteristic selection are preprocessing operations that raise the data standard. The metrics used in the data, such as glucose, blood pressure, skin thickness, insulin, and BMI, have no value and are impractical for exercise. These are therefore regarded as missing information, and the average score of the feature column that contains the missing value is reinstated. The database would include 268 cases of diabetes, 500 samples from the PIMA dataset that were not diabetic, and 520 instances from the more evenly distributed Banglasesh Dataset.

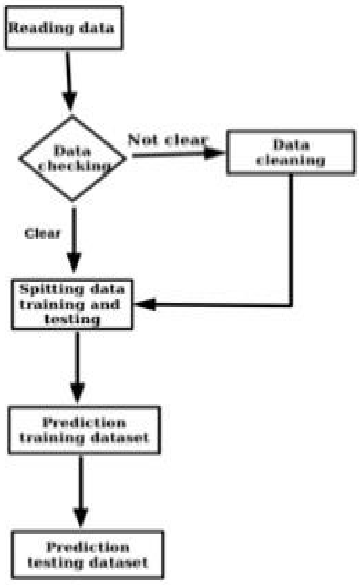
Reading data
Data checking
Not clear
Data cleaning
Clear
Spitting data training and testing
Prediction training dataset
Prediction testing dataset

Fig 2 Pre-processing the data

B. Splitting of Dataset

The train-test split methodology is used to evaluate classification algorithms using to extract results from data that is not used to train the model. After data initialization and rinsing, the sample is prepared for coaching and testing. It is common practise to evaluate using the average of all recorded ratings. The train-test split is a way for evaluating an ML algorithm's effectiveness. It can be used for classification or reversing works and is used for the classification algorithm. This study split our entire data set between 70% to 30% ratio for training data and testing data respectively.

C. Model Training

This is the most crucial step, which includes the This study presents the results of 10 classifiers of machine learning. A brief expression of these methods are given below:

1) Logistic Regression Classifier (LRC)

Classification and regression issues are amenable to being solved with the help of the supervised learning method known as logistic regression. Probabilistic models are developed to categorise categorical data [10, 11]. In order to make an accurate prediction of the result through a sigmoid function and distribute the input parameters in an equal manner. The probabilities of the sigmoid function is utilised to hypothesise the facts that are most likely to take place, and the probability might range from 0 to 1 to indicate whether or not an event takes place. The judgement bar transforms into a label for a class when it is used. This could be an example of binary (with either 0 or 1), multinomial (with three or more classes in any order), or ordinal data. It is easy to configure and has the capability of providing an accurate predictions.

$$P = 1/(1 + e^\wedge - (a + bX)) \qquad (1)$$

In Eq. (1), p stands for probability, a and b are elements of the model, and X is a factor.

2) K-Nearest Neighbour Classifier (KNN)

K-nearest neighbour is one technique for supervised classification [8]. It sorts objects into categories based on their proximity to one another. One example is situation-based instruction. It finds out the distance(Euclidian/Mhalonobi) among a feature and its closest neighbour. It utilises the efforts of numerous individuals. To determine how to name a new point, it examines a catalogue of previously named points. The facts are organised into categories based on their similarities. They can be used in conjunction with KNN to fill in absent digits. Many predictions were made after the undetermined numbers. The data set is processed by methodologies [1] to improve. If you utilise them differently, you may be able to increase your productivity. In this analysis, K=7 is the optimal K value.

3) Support Vector Machine Classifier(SVM)

SVM creates a hyperplane with the largest possible gap to separate various types of data or to retain similar points on either side. SVM can be used to identify and categorise objects. [11] A hyperplane can be used to predict cardiac disease by plotting the probability of having cardio disease on the x-axis and the probability of not having cardio disease on the y-axis. Image 1 depicts a straightforward Classifier curve. Two types of SVM exist: linear and nonlinear. When a linear line cannot distinguish between label data, non-linear does the work. Linear SVM is utilised to partition the data when no edge is required. Non-linear SVM employs a kernel function when standard SVM cannot work out how to split complicated data.

4) Naïve Bayes Classifier (NBC)

One of the most popular algorithms in ML is Bayes Theorem based naive Bayes classifier (also known as a hidden Markov model) used to classify continuous variables. It assumes that the probabilities of the parameters finds joint probability distribution over all feasible parameter values assuming independence of features to each other. Here, Bayes theorem is used to measure the posterior probability. These calculations are shown in the below equation:

$$P(m \mid y) = (P(y \mid m)p(m))/(P(y)) \qquad (2)$$

P(m|y) is posterior probability, while P(m) is called prior probability.

5) Decision Tree Classifier (DTC)

A decision tree is a method of probabilistic classification that may be applied to both category and numerical data. A structure that resembles a tree is referred to as a decision tree [4]. In the process of resolving issues involving patient records, decision trees are becoming increasingly prevalent and indispensable. The data in a tree graph is uncomplicated to generate and analyse. Trees base-nodes are required to construct an analysis using a decision tree model. This strategy organises clusters or form groups that are more closely related based on the most crucial indicators. The entropy of each attribute can be calculated as follows:

$$Entropy(X) = \sum_1^n (P(X_i) * LogP(X_i) \qquad (3)$$

6) Random Forest Classifier (RFC)

The random forest (RF) technique is a supervised classifier. In this procedure, multiple trees are used to create a forest. The model's prediction is based on the category that received the most votes. The more trees there are, the more accurate a random forest (RF) classifier is.

7) Gradient Boosting Classifier (GBC)

The objective of gradient boosting is to minimise the model's loss function. Using gradient descent iterative first-order optimisation, it is possible to determine the local minimum of a differentiable function. Because it is founded on the minimization of a loss function, gradient boosting is a versatile technique that can be used for regression, multi-class classification, and other purposes. Gradient boosting can also be utilised in other professions to improve accuracy.

8) Ada Boost Classifier (ABC)

Adaptive Boost is an ensemble technique that emphasises misclassified observations for training by modifying the sampling distribution to increase the weights of such error-prone points.

9) XgBoost Classifier (XBC)

XGBoost is also a gradient boosting category ensemble technique that utilises the histogram method to determine the optimal data bin division.

10) Extra Tree Classifier (ETC)

It is similar to the random forest classifier with the exception that it selects random split points in decision trees rather than a random sampling of trees to prevent overfitting and then averages the results of multiple subsamples to improve accuracy.

V. RESULTS AND DISCUSSION

We have experimented on two datasets using 10 ML algorithms. Table 3 and 4 are showing the results obtained from PIMA and BanglaDesh Dataset respectively for both of the classes negative and positive. This study's purpose is to see if an individual will acquire heart disease. In this study, the efficacy of supervised machine learning classification strategies such as LRC, KNN, SVM, NBC, DTC, RFC, GBC, ABC, XBC, etc. was evaluated by comparing their individual performance. With the aid of the SkLearn package, a number of experiments were carried out in which several methods for categorization were used. The study utilised a 6th generation Intel Corei3 processor with a 3300H CPU rated at up to 2.1 GHz and 4 gigabytes of RAM. The data are examined as soon as feasible to ensure the overall accuracy of the employed procedures.

If we compare the results obtained from the two datasets, It may be clearly shown that accuracy and F1-score mainly depends on the dataset. If input is highly linearly correlated with output, then humble logistic regression may give a better efficiency. Otherwise, to handle non linearity we may need to apply neural network for the same. From table 3, it can be deduced that LR method is giving better result for accuracy and negative class but for diagnosis or positive class RFC is performing better. Since the output is highly imbalanced in PIMA dataset, we must accept F1-score as the better measurement.

Table 4. Performance Comparison of different ML for PIMA data set

Sr. No.	**Model**	**Accuracy (%)**	**Negative Class F1-score(%)**	**Positive Class F1-score**
1	LRC	**82.46**	**88**	68
2	KNN	75.97	83	59
3	SVC	79.22	86	60
4	NBC	75.73	85	64
5	DTC	73.37	82	49
6	RFC	81.57	87	**70**
7	ABC	77.92	84	65
8	GBC	81.81	86	68
9	XBC	78.57	85	42
10	ETC	80.51	86	67

As the majority of the inputs, with the exception of age, are binary, the BD dataset produces a one-of-a-kind output, as can be seen in table 4. In this case, RFC and ETC have provided a result or diagnosis of diabetes in every single patient. For RBC as well, it's close to one hundred percent. Hence, when we think about the BD dataset, we might take into consideration these three techniques.

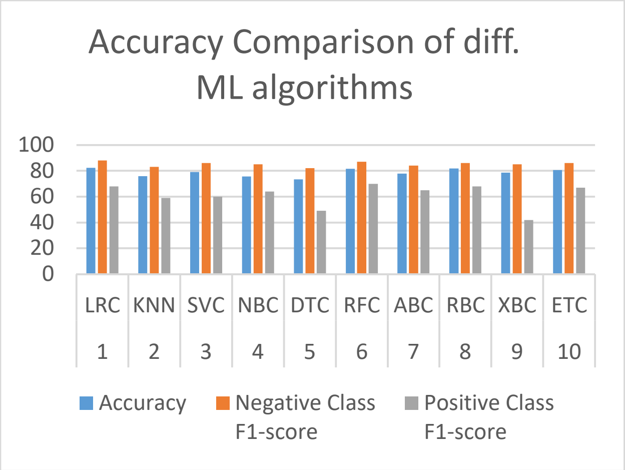

Fig 3. Accuracy and F1- comparison of diff. ML algorithms for PIMA dataset

Table 3. Performance Comparison of different ML for BD data set

Sr. No.	**Model**	**Accuracy**	**Negative class F1-Score**	**Positive class F1-Score**
1	LRC	86%	85%	88%
2	KNN	72%	71%	73%
3	SVC	77%	73%	77%
4	NBC	85%	84%	86%
5	DTC	95%	94%	95%
6	RFC	100%	100%	100%
7	ABC	89%	88%	90%
8	RBC	99%	99%	99%
9	XBC	88%	87%	88%
10	ETC	100%	100%	100%

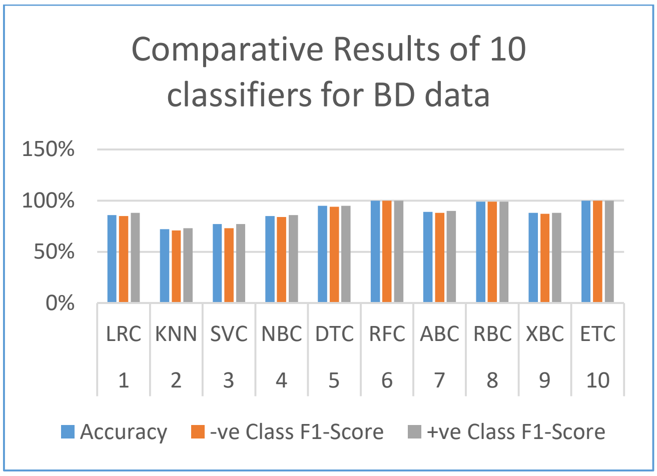

Fig 4. Accuracy and F1- comparison of diff. ML algorithms for BD dataset

Surprisingly, using PIMA dataset we found that simple linear regression is providing best result of 82% accuracy while on Bangladesh dataset ensemble base tree classifiers viz. RFC and ETC are providing the best result with 100% accuracy. This may be due to a smaller number of instances in Bangladesh dataset or less non-linearity in attributes. This research presented evidence that two distinct diabetes datasets each yield different approaches as the best way in terms of accuracy and F1-score, leading the authors to the conclusion that data modelling is highly dependent on data.

VI. Conclusion

This research comes to the conclusion that machine learning techniques can be utilized to detect diabetes quickly and correctly. This will allow for early identification of the disease, which will help in controlling the aftereffects of the disease through preventative measures done at the proper time. In addition, in comparison to conventional diagnostic establishments, this alternative would result in significant cost and time savings.

In this paper two different dataset namely PIMA and Bangladesh Dataset were used to establish the best machine learning algorithm. Surprisingly, using PIMA dataset we found that simple linear regression is providing best result of 82% accuracy while on Bangladesh dataset ensemble base tree classifiers viz. RFC and ETC are providing the best result with 100% accuracy. This may be due to less number of instances in Bangladesh datset or less non-linearity in attributes.

This research presented evidence that two distinct diabetes datasets each yield different approaches as the best way in terms of accuracy and F1-score, leading the authors to the conclusion that data modeling is highly dependent on data. Nonetheless, Random Forest is one of these algorithms that is producing superior results than other ways in both of the studies involving the datasets. So, based on this, we can get the conclusion that Random Forest is a more effective method of data classification.

References

[1] Global report on diabetes, 2016. Available at: https://apps.who.int/iris/bitstream/handle/10665/204871/9789241565257_eng.pdf;jsessionid=2BC28035503CFAFF295E70CFB4A0E1DF?Sequence=1

[2] Diabetes: Asia's 'silent killer, november 14, 2013. Available at: www.bbc.com/news/world- asia-247402

[3] Sarwar N et al. Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies. Lancet. 2010 Jun 26;375(9733):2215-22. doi: 10.1016/S0140-6736(10)60484-9.

[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] Srivastava, Suyash Sharma, Lokesh Sharma, Vijeta Kumar, Ajai Darbari, Hemant. (2019). Prediction of Diabetes Using Artificial Neural Network Approach: ICoEVCI 2018, India. 10.1007/978-981-13-1642-5_59.

[6] Adhi Tama, Bayu & Rhee, Kyung Hyune. (2019). Tree-based classifier ensembles for early detection method of diabetes: an exploratory study. Artificial Intelligence Review. 51. 355-370. 10.1007/s10462-017-9565-3.

[7] Krishnan, Vijiyakumar Lavanya, B. Nirmala, I. Caroline, S.. (2019). Random Forest Algorithm for the Prediction of Diabetes. 1-5. 10.1109/ICSCAN.2019.8878802..

[8] Anjali C, Veena Vijayan V, "Prediction and Diagnosis of Diabetes Mellitus -A Machine Learning Approach", IEEE Recent Advances in Intelligent Computational Systems (RAICS)-2015.

[9] Dr. Kayal Vizhi, Aman Dash, "Diabetes Prediction Using Machine Learning", International Journal of Advanced Science and Technology-2020.

[10] Jobeda Jamal Khanam, Simon Y. Foo, "A comparison of machine learning algorithms for diabetes prediction", The Korean Institute of Communications and Information Sciences (KICS)-2021.

[11] Islam, M. F., Ferdousi, R., Rahman, S., & Bushra, H. Y. (2020). Likelihood prediction of diabetes at early stage using data mining techniques. In *Computer Vision and Machine Intelligence in Medical Image Analysis: International Symposium, ISCMM 2019* (pp. 113-125). Springer Singapore.

[12] Schulz LO, Bennett PH, Ravussin E, Kidd JR, Kidd KK, Esparza J, Valencia ME (2006) Effects of traditional and western environments on prevalence of type 2 diabetes in Pima Indians in Mexico and the US. Diabetes Care 29(8):1866–1871

[13] Lejla Alic, Hasan T. Abbas, Marelyn Rios, Muhammad AbdulGhani, and Khalid Qaraqe, "Predicting Diabetes in Healthy Population through Machine Learning", IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)-2019.

[14] Ayman Mir, Sudhir N. Dhage, "Diabetes Disease Prediction using Machine Learning on Big Data of Healthcare", Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)-2018.

[15] Mitushi Soni, Dr. Sunita Varma, "Diabetes Prediction using Machine Learning Techniques", International Journal of Engineering Research & Technology (IJERT)-2020.

[16] Md. Kamrul Hasan, Md. Ashraful Alam, Dola Das, Eklas Hossain, Mahmudul Hasan, "Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers", IEEE-2020.

[17] H. Suriya Babu, T. TamilArasan, P. TheepakPrakash, "Prediction of Diabetes Mellitus Using Machine Learning Algorithms", International Journal of Advanced Engineering Science and Information Technology (IJAESIT)2021.

[18] Abdulhakim Salum Hassan, I. Malaserene, A. Anny Leema, Diabetes Mellitus Prediction using Classification Techniques, -International Journal of Innovative Technology and Exploring Engineering (IJITEE)-2020