PREVALENCE AND EARLY PREIDCTIONOF DIABETES USING MELLITUS CLASSIFICATION BASED ON MACHINE LEARNING ALGORITHM

[1]Mr.M.Kishore Varma, [2]Dr.L.V.Ramesh, [3]Dr. D. Prabhakar

[1] Assistant Professor, Dept of CSE, SRKR Engineering College (A), Chinnaamiram, Bhimavaram, West Godavari District, Andhra Pradesh, India
[2]Associate Professor, Dept of IT, CMR Technical Campus, Medchal Road, Kandlakoya, Hyderabad, Telangana, India
[3]Associate Professor, Dept of ECE, Seshadri Rao Gudlavalleru Engineering College, Krishna District, Andhra Pradesh, India
Mail Id's:-[1]kishoresrkrcse@gmail.com, [2]rameshlv.it@cmrtc.ac.in, [3]prabhakar.dudla@gmail.com

ABSTRACT: Throughout the world, diabetes is one of the biggest health complications, which is faced by so many people. The uncontrolled diabetes increases the risk of heart attack, cancer, kidney damage, blindness, and other illnesses, will increase highly. Type 2 (Mellitus) is considered as the most common disease in adults and it occurs when the body doesn't produce enough insulin or shows insulin resistance. The prevalence and number of diabetes cases have been increasing over the years. Hence a real time and accurate diagnosis is essential to prevent the diabetes and it can help to identify either a patient have diabetes or not. In recent times, the healthcare studies have used different advanced and innovative technologies to predict and diagnose the disease based on healthcare data. Machine learning is one of those techniques which are used to predict the disease more accurately. Hence in this work, Prevalence and Early Prediction of Diabetes using Mellitus Classification based on Machine Learning Algorithm is presented. In this system, three ML classifiers namely Random Forest (RF), ANN(Artificial Neural Network) and SVM(Support Vector Machine) are used. This system validates the performance of each algorithm in terms of Precision, Accuracy, Specificity and Sensitivity. This
approach achieves better results for the early prediction of diabetes.

KEYWORDS: Diabetes, Mellitus, Early Prediction, Machine Learning, Artificial Neural Network.

Doi: 10.31838/ecb/2023.12.8.809

I. INTRODUCTION

Diabetes is one of the chronic metabolic diseases which can be characterized by the increasing blood glucose levels due to the insufficient production of insulin. It is a non-communicable disease and it affects the person healthcare while reducing the efficiency of a person.

When compared to the normal glucose level in the body, the rise of blood glucose level is more in this disease. It is important to mention that, for better metabolism, glucose is needed to the body, which is in the form of sugar. All cells want glucose as a source of energy. Due to lack of insulin hormone in the body, the blood glucose level has been increasing. Then, it imbalances the blood glucose in the body and causes severe damage to other parts of the body, like kidneys, heart, eyes and so on [4].

To diagnose and predict the diabetes at an early stage is one of the challenging tasks. As a result, to predict the diabetes disease, an accurate and efficient method is needed. Diabetes is a chronic condition that spreads throughout the world. The diabetic patients are, asymptomatic and they don't know about their condition. Diabetes frequently goes undiagnosed because of the nature of the diseases. Some body parts like heart, kidneys, nerves, blood arteries, and eyes are suffering severe long-lasting damages because of this uncontrolled diabetes. As a result, early disease detection is needed to improve the quality of life of the patients who are at the risk to take the preventive procedures and to slow down the disease progression.

This is a chronic, metabolic disease which is classified by the increasing blood sugar levels (i.e. glucose) in human body. Mainly the diabetes can be classified as two types an they are, Type I and Type II.

One of the most general being is Type 2 which occurs due to resilience developed in the body against the insulin or its non-synthesis developed by the pancreas. Type I is a chronic disease which is recognized as an insulin dependent disease. When compared to normal person's standard level, the patient's sugar level is high in Pre diabetes. However, this is not to the level of Type II. In general, it doesn't show any symptoms but sometimes skin can get darkened up to some extent. Generally Gestational diabetes is occurred in ladies during pregnancy. Due to insulin expansion and variety of hormones, during pregnancy, in some ladies Glucose level gets augmented. To maintain the fasting blood glucose, random blood glucose, and glucose resilience levels, diabetes is analyzed [7].

One of the important tasks in the chronic illness like diabetes, is prediction of the probability of an individual's risk and susceptibility. We need to diagnose chronic illness at an early stage to reduce the risk of more complicated health problems and also to save on medical costs. It is pertinent that deductions can be made accurately cases with the help of immediately measurable medical indicators, to help clinicians for making better decisions to save the patient who is at high risk situations and even in emergency cases like, the patient may be in the position of unconsciousness or unintelligible [2].

Due to their vital progressions, both big data and bio-technology provides an immense ability to predict, diagnose, and find a cure for the disease [1]. Now-a-days, healthcare studies have applied different technologies to predict and diagnose the patient based on the collected clinical information, Recently, with the help of Machine Learning techniques, the healthcare system can predict diabetes more accurately.

Artificial Intelligence has been used to diagnose the disease for several years. For detecting various types of diseases, it provides outstanding outcomes in predicting the outbreak of pandemic in any region or country, and for many other applications. To find out diabetes at an early stage, a brilliant ML-based diagnostic method can be used, and it also gives us accurate results. For training purpose, a suitable dataset with relevant features is required. A ML-based system is needed, to identify the presence and absence of diabetes. Over the years a lot of research is going on diabetes prediction and diagnosis [3].

Machine learning enables a computer to become intelligent by learning from the inputs or experiences (i.e. healthcare data) and predict the disease. Various researches have been performed like disease detection, prediction and diagnosis. From the recent researches, it is clear that, Machine Learning (ML) algorithms have been used in disease identification and prediction [6].

Over the past years, prevalence and the cases of diabetes have been increased. An accurate and timely diagnosis will help the patients for the prevention of diabetes at an early stage and it helps the patients to know if they get diabetes or not. However, the medical resources are limited and the doctor only make diagnosis for certain number of patients in the limited time [9]., Most of the people makes assessment based on their symptoms and experience However, for patients, it is inaccurate to make diagnoses for themselves because most of the patients are lacking of knowledge about medical profession and

they are just based on what they know and what they hear. Hence, it is important to make an efficient prediction model that can able to diagnose themselves and save medical resources.

Hence this work presents Prevalence and Early Prediction of Diabetes using Mellitus Classification based on Machine Learning (ML) algorithm. The remaining work is arranged as follows: the different research works on prediction of diabetes, is described by section II. Prevalence and Early Prediction of Diabetes using Mellitus Classification based on Machine Learning Algorithm is described in section III. Result analysis of presented approach is evaluated section IV. And at the end, section V provides the conclusion.

II. LITERATURE SURVEY

Mafas Raheem et. al., [5] A Deep Neural Network and a Data Science Approach will help to Predict Diabetes. To improve the accuracy, Deep Neural Network model is used. The stages include pre-processing, class balancing, normalization, feature selection process and hyper-parameter tuning using the cross-validated searching technique. In diagnosing or detecting diabetes, the model is good and as a promising predictive model and it achieved 0.88 ROC score and 88% of accuracy.

Hala Alshamlan, Hind Bin Taleb, Areej Al Sahow et. al., [8] described A Gene Prediction Function by using Logistic Regression for Type 2 Diabetes Mellitus. To predict diabetic and non-diabetic people by processing two datasets, this approach is very useful. Logistic regression classifier is used for Feature selection. With the help of fisher score feature selection, Ch-2 feature selection obtained high accuracy for logistic regression compared to other algorithms. The results of accuracy showed 90.23% and 61.90% respectively for two datasets.

Sandeep Bhargava, Kamlesh Lakhwani, Kamal Kant Hiran, Mahesh M, et. al., [10] describes Prediction of the Onset of Diabetes Using Artificial Neural Network (ANN) and Pima Indians Diabetes Dataset. Here, an automatic diagnosis system introduced and analyzed. Pima Indians Diabetes dataset and a Three-Layered ANN are used. Logistic-activation function is used as an algorithm for Activation of neurons. The Quasi Newton method is used for the training. This model achieved maximum gain score and cumulative gain plot is achieved.

An Dinh, Stacey Miertschin, Amber Young and Somya D. Mohanty et. al., [13] describes A data-driven approach to predicting diabetes and cardiovascular disease with ML. Authors evaluated the ML models and their capabilities. In detecting at patient's risk, based on the survey data and laboratory results, it identifies the key variables within the data. For cardiovascular disease (based on 131 variables), an Area Under-Receiver Operating Characteristics, (AU-ROC) is achieved with the help of laboratory results by the developed ensemble model. Authors get the score of 83.9% accuracy using no laboratory results and achieved 83.1% while using laboratory results.

Muhammad Azeem Sarwar, Nasir Kamal, Wajeeha Hamid, Munam Ali Shah et. al., [15] presents Prediction of Diabetes Using Machine Learning Algorithms in Healthcare. This analysis discusses the predictive analytics in healthcare with six different types of machine learning algorithms. For experiment purpose, dataset of medical record of patients which is obtained from internet and six machine learning algorithms are applied. There is a comparison and discussion is made on applied algorithms performance. Different machine learning techniques performance is evaluated in this study. With the help of machine learning techniques, this approach will help the

doctors and practitioners in early prediction of the diabetes.

Messan Komi, J un Li, Y ongxin Zhai, Xianguo Zhang et. al., [16] describes Application of Data Mining Methods in Diabetes Prediction. Through five different data mining methods, early prediction of diabetes is explored. The data mining methods including: GMM (Gaussian mixture model), SVM (Support Vector Machine), Logistic regression, ELM (Extreme Learning Machine) and ANN (Artificial Neural Network). These models are trained and validated on the dataset. In response to the predictable states, all the five models are able to extract patterns. The experiment result proves that, ANN (Artificial Neural Network) provides better accuracy compared to other algorithms. To predict patient with diabetes, the most effective model appears to be ANN followed by ELM and GMM.

P. Suresh Kumar, S. Pranavi et. al., [17] describes Performance Analysis of Machine Learning Algorithms on Diabetes Dataset using Big Data Analytics. The target of this analysis is based on different metrics like sensitivity, specificity accuracy, kappa, precision, and recall, to identify, analyze and compare different machine learning algorithms. With the help of Random Forest (RF), SVM, K-NN, CART (Classification and Regression Tree) and LDA (Linear Discriminant Analysis) algorithms, an inclusive study is done on diabetes dataset. The achieved results showed that, when compared to other algorithms, RF is producing more accurate predictions.

Sreekanth Rallapalli, Suryakanthi T et. al., [19] describes Predicting the Risk of Diabetes in Big Data Electronic Health Records by using Scalable Random Forest Classification Algorithm.Authors described Scalable Random Forest classification algorithm, to describe a predictive model which can identify the classifier rate for risk of diabetes can be done accurately. Based on scalable random forest algorithm, this analysis determined the nearest accuracy of the classifier. The CART model and RF are used. Results show that the Scalable RF algorithm obtained the nearest accuracy of the prediction.

III. PREVALENCE AND EARLY PREDICTION OF DIABETES

In this section, Prevalence and Early Prediction of Diabetes using Mellitus Classification based on Machine Learning Algorithm is presented. The block diagram of presented approach is shown in Fig. 1.

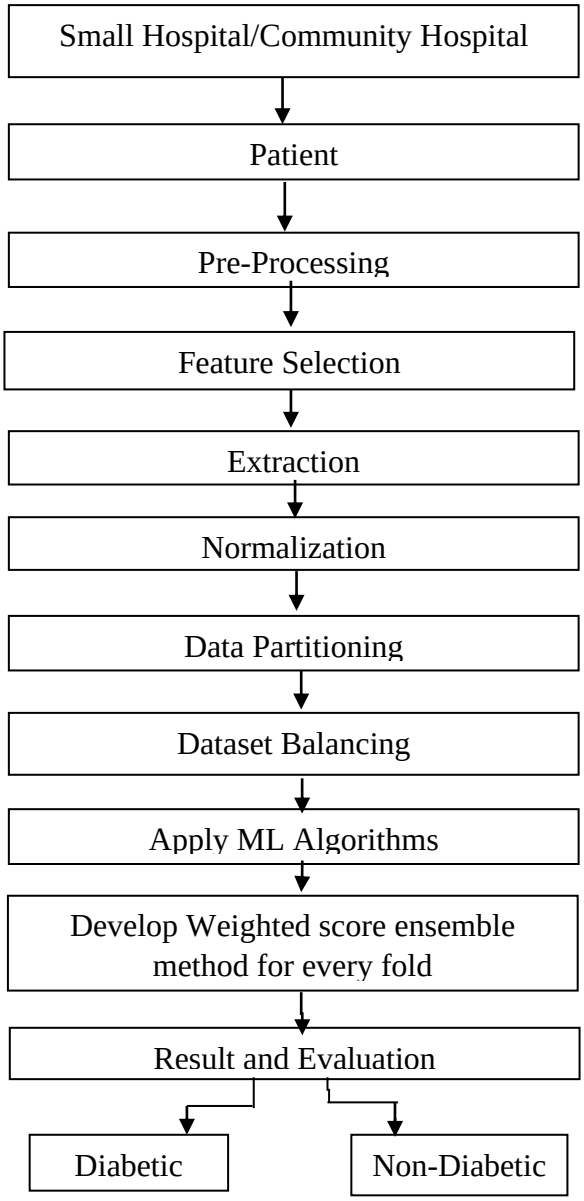

Fig. 1: Block Diagram of Prevalence and Prediction of Diabetes at an early stage

The prediction of an individual probability and susceptibility to a chronic disease like diabetes is one of the important tasks. Diagnosing the chronic illness at an early stage reduces the risk of more complicated health problems and saves on medical costs. To help clinicians make better decisions for patient treatment in high-risk situations, it is pertinent that deductions can be made accurately from immediately measurable medical indicators, even in emergencies where a patient may be unconscious or unintelligible.

A community hospital can be purely a nominal designation or have a more specific meaning. When specific, it refers to a hospital that is accessible to the general public, there a general or specific medical care will be provided, which is usually short-term, in a cost-effective setting, not only focuses on preventing illnesses but also treating them. In the name of hospital, the word community often occurs. For a community hospital, the word community is used in the sense of a location-based community. A patient is any recipient of health care services, the patient details are collected at the hospital, like age, name, BMI (Body Mass Index), Blood Pressure, Glucose, Insulin level, etc. These all things are performed by healthcare professionals.

Here, the dataset of Pima Indian diabetes is used. This dataset is originated, from the National Institute of Diabetes and Digestive and Kidney Diseases. To diagnostically predict whether a patient has diabetes or not, this is the objective of the dataset and this is based on certain diagnostic measurements included in the dataset. On the selection of these instances from a larger database, several constraints were placed. Here all patients particularly, females at least 21 years old of Pima Indian heritage. One target variable outcome and several medical predictor variables are consisted, by the datasets. Predictor variables include the number of pregnancies that the patient has, their BMI, insulin level, age, and so on. The following attributes are considered, in order to decide a female in this dataset as diabetic.

Glucose: It shows the level of plasma glucose concentration (2 hours).
Age: It exhibits the age in years.
Pregnancies: It indicates the number of times a female can get pregnancy.
Blood pressure: It shows the diastolic blood pressure in mm Hg (millimeters of mercury)
Skin thickness: It shows the triceps skin thickness in mm.
Insulin: Its range is in between 0 to 846. The average is 80.
BMI: Body Mass Index has been shown in Kg/m2.
Diabetes pedigree function: The likelihood of diabetes scored by this function.
Outcome: It is either 0 or 1. Here the meaning of 0 is that, a female has non-diabetic and the meaning of 1 is that, a female is diabetes.

In this dataset, to predict that the Pima Indians female is diabetic or not, the outcome used as a target class. It has 768 instances (i.e., number of rows) and 9 columns (i.e., number of attributes). When they are not equally balanced, classifications make prediction modelling more difficult. Data Pre-processing technique is used, to convert the raw data into a clean data. From the applied model, for achieving better results in Machine Learning projects, the format of the data has to be in a proper manner. From the database, in Pre-processing stage, undecipherable values known as, standard formatting and errors in data types are converted to null representations. In the feature extraction phase, the patient records are represented as a class label and features of data frame. In building a

promising predictive model, Feature selection is another important process.
With the help of most significant variables, it is better to check and select the input variables, although the variables seem to be important for the prediction of the chosen target. The features are like an array, towards the patient information collected from the laboratory, demographic, and survey methods. While preserving the information in the original data set, the transformation of raw data into numerical features can process. This is referred by Feature extraction.

In the raw data, there are two main problems. In that one is Dominant features and other one is outliers which could put a stop to the learning of machine learning algorithms. To obtain uniform contributions from each feature, Normalization is performed on the data by using the following standardization model. The data normalization can either transforms or rescales the raw data.

$$x' = \frac{x - \bar{x}}{\sigma} \quad (1)$$

Where x is considered as the original feature vector, the mean of that feature vector is considered as $\bar{x}$, and its standard deviation is considered as σ. After the entire dataset is stratified, to the dataset, a five-fold cross-validation is applied.

Usually, the dataset contains noise, missing values, outliers and so on. Towards these noises, the final result of the model may mislead by predictive machine learning algorithms because, they are sensitive. Therefore, before building the most profitable predictive model, cleaning of data and suitable strategies should be performed. Further, class balance must be performed in the target variable for overcoming the biases in model building.

As per the dataset, missing values are not found. However, in certain variables, zero values could not be accepted. Due to this reason, those values were replaced by missing.
Blood Pressure: 0 blood pressure is not found in living persons.
Glucose Level: 0 glucose level is not found even in fasting people.
Thickness of skin: Thickness of skin can't be less than 10mm, when it considered to normal people.
BMI: Unless in serious sick, it should not be 0.
Insulin: Having 0 insulin, happens in very rare cases.

Based on general biological knowledge, the reasons are drafted. By using median values, the missing values are then imputed. To improve ML models performance, the data normalization or scaling is always expected. Several methods being used in this phase, and to make a dataset balanced, the input variables are also normalized. In this work, the SMOTE (Synthetic Minority Over-sampling Technique) Tomek is applied. SMOTE Tomek is a combination of the SMOTE and the Tomek algorithms. For synthetic minority oversampling technique, the SMOTE is an acronym. Tomek is called as under sampling technique. At first, to create new synthetic minority samples and to get a balanced distribution of the classes, SMOTE will be applied.

Moreover, separation between the two classes will increase by using the link called Tomek. This link also useful for removal of the samples that are close to the boundary of the two classes. The test set remained the same for the train dataset, and it is applicable for the same. By scaling the data set, all values can be balanced. Random Forest, Support Vector Machine (SVM) and Artificial Neural Network (ANN) are the three different ML models. Next balanced data is applied to ML models for diabetes prediction. These are used and described in detail as follows:

The RF is a boosting technique. Wherever, base learners will generate, the parallel ensemble methods are applied in parallel. With a sample of data, each base learner model is provided. And these base learners will give the result separately. At last, based on the voting classifier the final prediction is made with the help of base learners. To obtain a more accurate and stable prediction, the RF builds multiple decision trees and merges them. In the RF, the decision trees are the base learner models. Since, the errors are often reduced dramatically, by calculating the basic motivation of the parallel methods, and it is also useful for independence among the base learners.

Support vector machine is a linear model. For classification and regression problems, SVM is used. This is a linear model. SVM also allows, data reduction. With the help of the support vectors and class corner points, the marginal distance between the classes is calculated from the centre of the hyperplane. Kernels, coefficients, and intercepts are the some of the variables used in SVM. In the whole SVM, Kernel is the most crucial component. These kernels have been adjusted, with the help of data that they receive. The data is linear to RBF (Radial Basis Function) and it justifies the use of linear and Gaussian kernels in study. By using it, linear and at the same time nonlinear problems can be solved. By generating a hyperplane, the new data is classified by SVM, with a maximum marginal distance. It works like a linear regression.

Three main layers are there in ANN. In that first one is input, second one is hidden, and the third one is output layers. In the input layer, input will be given and from the output layer, output will be received. To adjust the weight and to reduce the error between the true value and the target value, the middle layers are used. This process is called as back-propagation. In the input layer, there are eight nodes and two middle layers contains ten and eight nodes respectively, in this ANN. In the output layer, there is only one node, because it is a binary classification. The usage of twenty epochs was done for each cross-validation and first two layers are used and for the last layer, the sigmoid is used.

The hyper parameters of algorithm should be tuned for any dataset to obtain the best result. To achieve the desired outcome, the hyper parameters of the algorithms are tuned. To select the best performing weights for the ensemble model, some weights must be evaluated and then the weights are selected, those weights will produce the best accuracy. Among the three algorithms, for ensemble method, the best two performers are used.

A loop is used to produce the highest accuracy for every fold and to check the combination of weights. A weighted model is developed, for the ensemble classification. For two algorithms, two weights are assigned. They used in the ensemble approach. Finally, if the patient has diabetes, then it is classified as type1 or type 2 Diabetes. This ensemble approach helps to predict the patient as diabetic or non-diabetic.

IV. RESULT ANALYSIS

In this section, Prevalence and Early Prediction of Diabetes using Mellitus Classification based on Machine Learning Algorithm is implemented using python. In this analysis, the dataset of Pima Indian diabetes is used. In this approach, the three ML algorithms are used. They are Random Forest, ANN and SVM algorithms, by using the confusion matrix parameters TP (True Positive), True Negative (TN), FP (False Positive) and False Negative (FN), the performance of these three algorithms is evaluated. These are the instances used to predict the diabetes. The performance of presented approach is evaluated, in terms

of Accuracy, Specificity, Sensitivity and Precision which, are as follows:

Accuracy: Accuracy is described in the form of ratio of correctly predicted instances to the total predicted instances and is expressed as

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \times 100 \quad (2)$$

Specificity: The percentage of correctly predicted Negative instances to the total Negative Instances is called as Specificity and is given as

$$Specificity = \frac{TN}{TN + FP} \times 100 \quad (3)$$

Sensitivity: It is also called as True Positive Rate (TPR). Percentage of correctly predicted Positive instances to the total positive instances is called as Sensitivity. And it is expressed as

$$Sensitivity = \frac{TP}{TP + FN} \times 100 \quad (4)$$

Precision: It is defined as the ratio of True Positive instances to the total Positive Predicted instances and is given as

$$Precision = \frac{TP}{TP + FP} \times 100 \quad (5)$$

The table 1 shows the performance comparison of three ML algorithms in order to determine the better algorithms for diabetes prediction and classification.

Table 1: ML Algorithms Performance Comparison

ML Algorithms	Accuracy (%)
RF (Random Forest)	89.78%
SVM (Support Vector Machine)	93.45%
ANN (Artificial Neural Network)	95.67%

Among these three ML algorithms, the ANN and SVM have better results. That is the reason, these two (ANN+SVM) are used with ensemble classification for diabetes prediction and classification. Finally, this approach predicted the diabetic and non-diabetic patients. In addition, it classifies the diabetes mellitus as type 1 and type 2. The overall performance evaluation of presented approach is evaluated in table 2.

Table 2: Performance Metrics Evaluation

Performance metrics	For Prediction of Diabetes, (DNN) Deep Neural Network: A Data Science Approach	Presented Prevalence and using mellitus classification for the Prediction of diabetes, based on Machine Learning Algorithm
Accuracy (%)	88.00	95.64
Precision (%)	88	94.56
Sensitivity (%)	90	95.23
Specificity (%)	89.32	94.23

The diabetes prediction performance of different approaches is compared in terms of Accuracy, Specificity, Sensitivity and Precision. Compared to DNN approach, presented approach based on ML algorithms has high accuracy, precision, sensitivity and specificity. The Fig.2 shows the graphical comparison in terms of sensitivity and specificity.

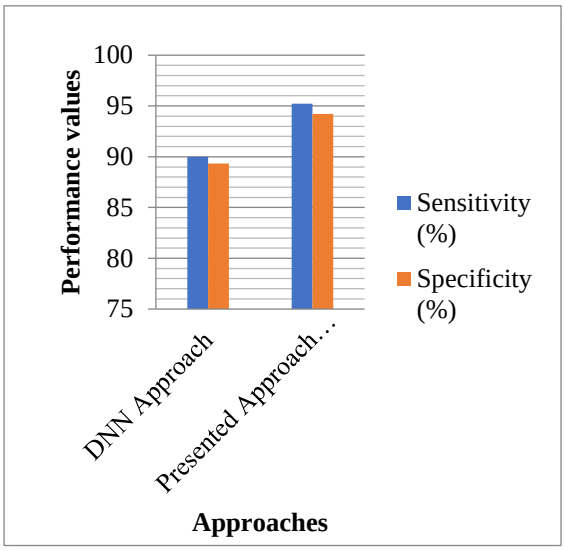

Fig. 2: Sensitivity and Specificity Metrics Comparison

algorithm has better results than DNN approach for diabetes prediction. The Fig. 3 shows the performance comparison in terms of accuracy and precision. In fig. 3, the x-axis represents different approaches whereas y-axis represents performance values.

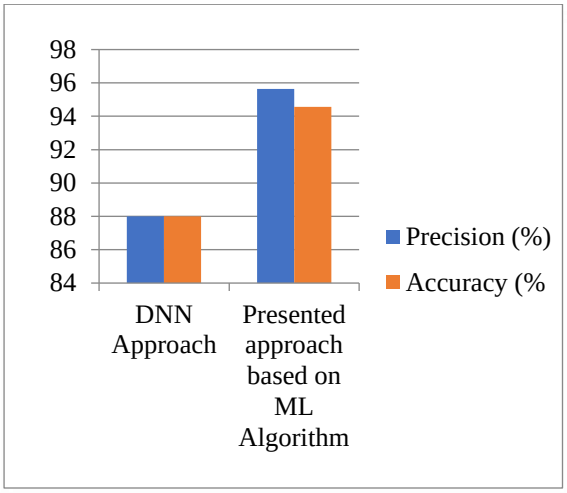

Fig. 3: Performance Comparison

Compared to earlier DNN approach, presented approach has higher Accuracy as well as higher Precision. Hence, presented approach is used a better solution for diabetes disease early prediction.

V. CONCLUSION

In this work, Prevalence and Early Prediction of Diabetes using Mellitus Classification based on Machine Learning Algorithm is presented. This approach is very useful for the prediction, whether a person has diabetes or not. If the person has diabetes, then it classifies it as type 1 or type 2 mellitus. Pima Indian diabetes dataset contains different attributes And this is used in this analysis. Different phases such as data pre-processing, feature selection, extraction, normalization and dataset balancing are performed to improve the performance of presented approach. Different ML algorithms namely RF, SVM and ANN are used here. In order to predict the diabetes in early stage, this approach uses ANN and SVM with ensemble method. Presented approach's performance is evaluated. The performance metrics of presented approach are compared with the earlier approach, to determine the effectiveness of presented approach, in terms of Precision, Accuracy, Specificity and Sensitivity. However, presented approach has better results than earlier approach. Therefore, for early diabetes prediction and classification, presented approach will be an accurate and effective solution.

VI. REFERENCES

[1] Mahendra Kumar Gourisaria, Gaurav Jee, G. M. Harshvardhan1 Vijander Singh, Pradeep Kumar Singh Tewabe Chekole Workneh, "Data science appositeness in diabetes mellitus diagnosis for healthcare systems of developing nations", IET Commun. 2022; 16:532–547, DOI: 10.1049/cmu2.12338

[2] Victor Chang, Jozeene Bailey, Qianwen Ariel Xu, Zhili Sun, "Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms", Neural Computing and Applications, Sringer, 2022, doi:10.1007/s00521-022-07049-z

[3] Hafsa Binte Kibria, Md Nahiduzzaman, Md. Omaer Faruq Goni, Mominul Ahsan and Julfikar Haider, "An Ensemble Approach for the Prediction of Diabetes Mellitus Using a Soft Voting Classifier with an Explainable AI",

Sensors 2022, 22, 7268, doi:10.3390/s22197268

[4] Sourav Kumar Bhoi, Sanjaya Kumar Panda, Kalyan Kumar Jena, P. Anshuman Abhisekh, Kshira Sagar Sahood, Najm Us Sama, Shweta Supriya Pradhan, Rashmi Ranjan Sahoo, "Prediction of Diabetes in Females of Pima Indian Heritage: A Complete Supervised Learning Approach", Turkish Journal of Computer and Mathematics Education, Vol.12 No.10 (2021), 3074-3084

[5] Mafas Raheem, "Deep Neural Network to Predict Diabetes: A Data Science Approach", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-9 Issue-6, March 2021,DOI:10.35940/ijrte.E5255.039621

[6] Saloni Kumari, Deepika Kumar, Mamta Mittal, "An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier", International Journal of Cognitive Computing in Engineering 2 (2021) 40–46, doi: 10.1016/j.ijcce.2021.01.001

[7] Mohd Javeed Mehdi, N Srinivasrao, A Sireesha, "Detection and prognosis of diabetes based on data science techniques", 2020 Elsevier, doi: 10.1016/j.matpr.2020.08.386

[8] Hala Alshamlan, Hind Bin Taleb, Areej Al Sahow, "A Gene Prediction Function for Type 2 Diabetes Mellitus using Logistic Regression", 2020 11th International Conference on Information and Communication Systems (ICICS), Doi: 10.1109/ICICS49469.2020.239549

[9] Juncheng Ma, "Machine Learning in Predicting Diabetes in the Early Stage", 2020 2nd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI), DOI: 10.1109/MLBDBI51377.2020.00037

[10] Kamlesh Lakhwani, Sandeep Bhargava, Kamal Kant Hiran, Mahesh M. Bundele, Devendra Somwanshi, "Prediction of the Onset of Diabetes Using Artificial Neural Network and Pima Indians Diabetes Dataset", 5th IEEE International Conference on Recent Advances and Innovations in Engineering-ICRAIE 2020 (IEEE Record#51050), DOI:10.1109/ICRAIE51050.2020.9358308

[11] Srinivasa Rao Swarna, Sumati Boyapati, Pooja Dixit, Rashmi Agrawal, "Diabetes prediction by using Big Data Tool and Machine Learning Approaches", Proceedings of the Third International Conference on Intelligent Sustainable Systems [ICISS 2020], ISBN: 978-1-7281-7089-3, DOI: 10.1109/ICISS49785.2020.9315866

[12] Ayman Alahmar, Emad A. Mohammed, Rachid Benlamri, "Application of Data Mining Techniques to Predict the Length of Stay of Hospitalized Patients with Diabetes", 2018 4th International Conference on Big Data Innovations and Applications, 978-1-5386-7793-3/18, 2018 IEEE,DOI 10.1109/Innovate-Data.2018.00013

[13] An Dinh, Stacey Miertschin, Amber Young and Somya D. Mohanty, "A data-driven approach to predicting diabetes and cardiovascular disease with machine learning", BMC Medical Informatics and Decision Making (2019) 19:211, doi:12911-019-0918-5

[14] K.VijiyaKumar, "Random Forest Algorithm for the Prediction of Diabetes", Proceeding of International Conference on Systems Computation Automation and Networking 2019, 978-1-7281-1524-5, 2019 IEEE

[15] Muhammad Azeem Sarwar, Nasir Kamal, Wajeeha Hamid, Munam Ali Shah, "Prediction of Diabetes Using Machine Learning Algorithms in Healthcare", Proceedings of the 24th International Conference on Automation & Computing, Newcastle University, Newcastle upon Tyne, UK, 6-7 September 2018.

[16] Messan Komi, J un Li, Y ongxin Zhai, Xianguo Zhang, "Application of Data Mining Methods in Diabetes Prediction", 2017 2nd International Conference on Image, Vision and

Computing, 978-1-5090-6238-6/17, 2017 IEEE.

[17] P. Suresh Kumar, S. Pranavi, "Performance Analysis of Machine Learning Algorithms on Diabetes Dataset using Big Data Analytics", 2017 International Conference on Infocom Technologies and Unmanned Systems (ICTUS'2017), Dec. 18-20, 2017, ADET, Amity University Dubai, UAE

[18] Ihsan Salman Jasim, Adil Deniz Duru, Khalid Shaker, Baraa M. Abed, Hadeel M. Saleh, "Evaluation and Measuring Classifiers of Diabetes Diseases", ICET2017, Antalya, Turkey, 978-1-5386-1949-0/17

[19] Sreekanth Rallapalli, Suryakanthi T, "Predicting the Risk of Diabetes in Big Data Electronic Health Records by using Scalable Random Forest Classification Algorithm", 2016 International Conference on Advances in Computing and Communication Engineering (ICACCE), DOI: 10.1109/ICACCE.2016.8073762

[20] Lin Li, "Diagnosis of Diabetes using a Weight-Adjusted Voting Approach", 2014 IEEE 14th International Conference on Bioinformatics and Bioengineering, 978-1-4799-7502-0/14 DOI 10.1109/BIBE.2014.27