2021 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT) | 978-1-6654-2306-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/CSNT51715.2021.9509732

A Neural Network based Diabetes Prediction on Imbalanced Data

Shivani Yadav
School of Information Technology, University Teaching Department,Rajiv Gandhi Technological University, Bhopal, Madhya Pradesh (462033) shivani12yadav@gmail.com

Yogendra P.S.Maravi
School of Information Technology, University Teaching Department,Rajiv Gandhi Technological University, Bhopal, Madhya Pradesh (462033) yogendra.rgpv@gmail.com

Jitendra Agrawal
*School of Information Technology, University Teaching Department,*Rajiv Gandhi Technological University, Bhopal, Madhya Pradesh (462033) jitendra@rgtu.net

Nishchol Mishra
School of Information Technology, University Teaching Department,Rajiv Gandhi Technological University, Bhopal, Madhya Pradesh (462033)
nishchol@rgtu.net

Abstract **As an extensively well-known chronic disease, diabetes is an illness that harms the body's capability to process blood glucose. The proper treatment of diabetes could help a person live a long and normal life in general. It is necessary to detect the disease at an early stage. We focus our work on the performance of a machine-learning (ML) algorithm to identify the presence of diabetes on the PIMA Indian diabetes dataset (PIDD) which referenced from the University of California, Irvine (UCI) ML repository. Using ML, we know about the classification and prediction techniques. Further, diabetes became an attention seeker in the field of research due to the presence ofimbalanced and missing data. Although many factors affect the performance of the algorithm, This research paper worked on the prediction technique for diabetes classification with outliers and missing values in data with class imbalance. Using an adaptive synthetic sampling method (ADASYN) and reduced the impact of class imbalance on the performance of the prediction model. Then, this algorithm improved the generalization using a feature selection technique and multilayer perceptron classifiers to make predictions and evaluations. Experimental results shows that this experiment obtained a better accuracy of 84% with a neural network model in comparison with the previous model.**

Keywords- Diabetes prediction**.** Machine Learning**.** Outliers**.** Artificial Neural Network**.** Adaptive synthetic sampling**.** Multilayer Perceptron

I. Introduction

Diabetes is the abbreviated version of full name diabetes mellitus. The term diabetes mellitus is derived from the Greek word diabetes which means siphon. The word 'siphon' means to pass through and the Latin word Mellitus means sweet [1]. This is reason when excessive sugar is found in the blood as well as the urine in the diabetic patient. In the 17th century, diabetes mellitus also known as pissing evil [2]. Diabetes is one of the most chronic diseases in the world. It was characterized by high blood sugar. It had become a fifth-ranked among various diseases for disease-related deaths [3]. Due to diabetes, other problems may arise like the increased risk of heart attacked and stroked, kidney failure, etc. The only way to protect from this disease was by managing the blood glucose level. Otherwise, this disease cannot be cured. Diabetes disease had mainly three categories such as type 1 diabetes, type2 diabetes, and gestational diabetes [4]. Other categories of diabetes mellitus were diabetic retinopathy and diabetic neuropathy. In type 1 diabetes, the condition depends on the occurrence of insulin mainly in children and teenagers. A genetic disorder was the main cause of this type1 diabetes. Other categories of diabetes mellitus were diabetic retinopathy and diabetic neuropathy. In type 1 diabetes, the condition depends on the occurrence of insulin pre-dominantly in children and teenagers. A genetic disorder is the leading cause of this type1 diabetes. These types of diabetic patients are also known as insulin-dependent diabetes mellitus. In type 2 diabetes, the detection of high sugar levels in the blood. It's mainly arising in adults during the age of 40years. Some factors are responsible for the cause of diabetes as a combination of genetic susceptibility, obesity, irregular food intake timings. The PIMA Indian dataset (PIDD) [5] was used for the development of prediction models for classified data into diabetic or non-diabetic based on their symptoms and through the performance of various algorithms applied on this dataset. Machine learning techniques had a massive perspective to improve the performance of the prediction model for diabetes. In terms of the development of the healthcare industry, they create a large amount of valuable data such as patient records, electronic medical history, the record of diagnosis, treatment data, etc. These patients related case history working as an indispensable source to extract knowledge that could help us to make a decision and also in reduction of cost. In general,

978-1-6654-2306-9/21/$31.00 ©2021 IEEE
DOI: 10.1109/CSNT.2021.89

data mining algorithms and techniques of ML have gained strength due to the potential of managing the record to a large extent for extracting knowledge, decision making, and making predictions.

II. RELATED WORKS

Various researched works implemented for the classification of diabetes. These considered some features for the diagnosis and treatment of diabetes. S. S. Dileep[6]works on the creation of an efficient model to compare various machine learning techniques using support vector machine (SVM), decision tree (DT), and Naïve Bayes (NB) to forecast the possibility of diabetes with maximum accuracy. They showed that the NB classifier performed best as compared to other methods with AUC 0. 809. A model based on ML techniques [7], where authors analyze the performance of different algorithms such as SVM, NB, k-nearest neighbor (KNN), DT, on adult people to predict diabetes. They demonstrate that DT achieved higher accuracy in comparison with other techniques employed in this model. The ML framework proposed [8], using linear discriminant analysis, quadratic discriminant analysis, NB, Gaussian process classification, AdaBoost (AB), logistic regression (LR) [9], SVM [10], ANN [11], DT [12], and RF [13] with dimension reduction and cross-validation method. They also worked on outlier rejection and filling missing values for boosting the performance of the ML model, where they chose AUC as a performance metric. Authors in [14], they developed the five different models to predict diabetes using a linear kernel SVM (SVM-linear), radial basis kernel, KNN [15], ANN, multi-dimensionality reduction methods. They also employed a feature selection method to avoid biased selection of dominant features in the dataset. The Boruta wrapper algorithm used for feature engineering. The proposed framework [16] performed the analysis of features in PIDD. Based on correlation values, they selected the optimal features from the dataset. They showed in their experiment, the feature selection method assist in improving the mapping of features effectively from low dimensions to high dimensions and provided the best fit to the data in aspects of diabetic or non-diabetic patients' records. In this experiment, many techniques such as DT, SVM, RF, NB used, where SVM gives the best results with improved accuracy. The artificial neural network (ANN) based classification model was proposed, in [17], they screened the PIDD data and worked with self-adaptive ANN. The neural network acclaimed by a cascading correlation (CC) algorithm. They trained the neural network using forward propagation and backpropagation to make classification easy. They demonstrated the CC with ANN helps to reduce the complexity of the neural network and also showed improvement in the performance of the technique. The deep learning technique [18] used to predict diabetes; they used a combination of a prevalent method from deep learning as long short-term memory technique and convolution neural network method. Then, it compared with the performance of the MLP classifier to show the impact of the technique. The predictive modeling [19] used to analyze the performance of different ML techniques based on a patient's record to help doctors in the early detection and treatment of diabetic patients. They performed on the PIDD data using different algorithms like KNN, LR, and gradient boosting, SVM, RF, DT [20], and neural network. In this experiment, RF performed better on this dataset with 79% accuracy.

In the previous work, we concluded that class imbalance plays a vital role in the classification of a dataset. To inscribe this challenge, we focus on improvement in the performance of the model to classify diabetes data. The presence of missing values and imbalanced class in the Pima Indian dataset had an enormous impact on the performance of the classifier. In this work, we were considering several factors that help to improve the accuracy of the technique.

III. MATERIAL AND METHODOLOGY

A. DATASET

The proposed work performed using the PIMA Indian dataset taken from the UCI Repository [5]. In medical data, missing value and class distribution is an unbalanced problem is common. The PIDD dataset consists of 768 instances with 8attributes, only 113 data without missing values. This dataset also has diabetic and non-diabetic samples. There are 268 diabetic patients and 500 non-diabetic patients, which presents a class imbalance in the dataset. It includes 8 numerical value features such as pregnant count, glucose concentration, blood pressure (mm Hg), Skin Thickness (mm), insulin (mm U/ml), body mass index (BMI), Diabetes pedigree function (PDF), and Age (years) as represents in Fig.1.

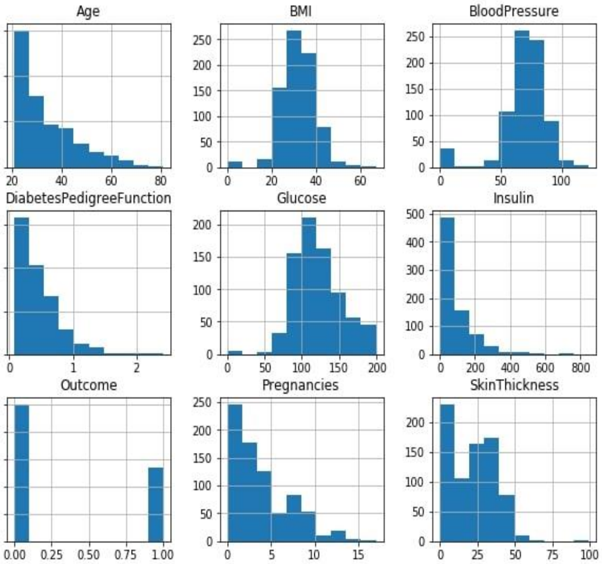

Fig. 1 Distribution of Attributes in PIDD

It comprises a class label, where the '0' represents the negative value of diabetes, and '1' represents a positive value of diabetes. The detailed description of the diabetes dataset mentioned in Table 1.

Table 1 Description of PIDD

Attribut e number	Attribute name	Description
1	Pregnancies	Number of times pregnant
2	Glucose	Plasma Glucose concentration
3	Blood pressure	Diastolic blood pressure(mm Hg)
4	Skin thickness	Skinfold thickness(mm)
5	Insulin	2-hour serum insulin(mu U/ml)
6	BMI	Body mass index(kg/m2)
7	Diabetespedigreefunction	Diabetes pedigree function
8	Age	Age in years
9	Outcome	Class label('0' or'1')

B. PACKAGES USED IN THE EXPERIMENT

Scikit-learn came filled with various features. Here some packages [21] were used in this experiment are as follows-

1. Numpy This library used for providing fast and effective operations with arrays. It is a modular extension of python. It is a package used for scientific computing and also handles a large number of arrays of homogenous data.
2. Pandas This package has a vital data structure in its dataframe. It built on a Numpy package. Data frames give opportunities for users to store and manipulate data in tabular form. This package opened the source. It is a powerful package for analyzing data. We have used this library to load the dataset as Pandas dataframe and analyzed data.
3. Matplotlib This package used for plotting 2d graphics in the python programming environment. It used in graphical user interface tool-kits. It is a type of plotting library as Matplotlib.pyplot for plotting the data. In an experiment, we have used this library to plot box-plot, a histogram of the dataset.
4. Seaborn This package used for statistical data visualization in a python programming environment. It based on the Matplotlib library. We have used this package to provide visualization of the dataset through a scatter plot.
5. Model Selection This package is a part of the Scikit-learn library in python programming. I imported train_test_split. So, we split the dataset into training and testing dataset. And import cross_valid_predict to check the accuracy of the prediction for the classification model.
6. Imbalanced-learn It is a python package that used for providing various re-sampling techniques. Here, we used adaptive synthetic oversampling methods for this experiment. It commonly used in the dataset. They were re-sampling techniques for showing the dataset strong as the imbalance-class was present in a dataset.

IV. PROPOSED FRAMEWORK

The problems faced by the prediction of diabetes can group in some aspects. First thing, the hurdle of missing values, the second one is the presence of outliers affects the performance of the prediction model, the third one is the problem of imbalanced class distribution. And scaling of the dataset also affects the overall performance of the classifier. So, the proposed framework designed to provide some optimal solutions to solve these issues. In proposed work, some preprocessing steps are composed as detection of outliers and replacing the value of outliers, filling missing values, features scaling, and feature selection of attributes.

The box-plot is the best method for visualizing the outliers in Fig. 2 and it uses an interquartile(IQR) method to show the presence of outliers. According to the range of IQR, replaces the value of outliers with the median value of attributes in the dataset instead of the removal of outliers shows in Fig. 3. The whole process mentioned in **Algorithm 1**. It improves the feature selection of attributes and shows some positive impact on the performance of the classifiers. The mathematical formulation for detecting and handling the outliers in this work is written in (1)

$$\text{Outliers}=\begin{cases} median(x), & if\ Lr \le x \le Ur \\ x\ , & Otherwise \end{cases} \quad (1)$$

where, x is feature value that lies on n-dimension space, Lr, Ur is the lower bound and upper bound of the attributes which are calculated through IQR formulation as written in (2), (3), and (4).

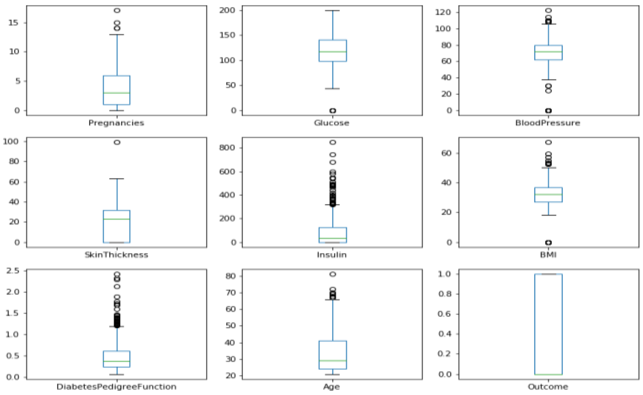

Fig. 2 Presence of outliers in PIDD

$$IQR = Q3 - Q1 \quad (2)$$

$$Lr = Q1 - (1.5 \times IQR) \quad (3)$$

$$Ur = Q3 – (1.5 \ \times IQR) \quad (4)$$

where, Q1, Q3, IQR is the first quartile, third quartile, inter-quartile range of attributes respectively

.

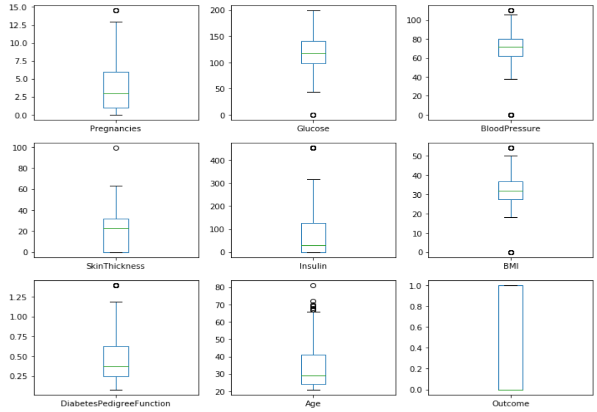

Fig. 3 After replacement of outliers in PIDD

Algorithm 1: The step to detect and handle the outliers using IQR method

```
1. Arrange dataset in increasing order
   sorted (df)
2. Calculate first quartile value (Q 1) and third quartile value (Q 3)
   Q1= df ['feature'].quantile (0.25)
   Q3= df ['feature'].quantile (0.75)
3. Find inter-quartile range(IQR)
   IQR=Q3-Q1
4. Find lower-bound(Lr)
    Lr = Q1 - (1.5*IQR)
5. Find upper-bound(Ur)
    Ur = Q3 + (1.5*IQR)
6. Anything that lies outside of lower and upper bound is an outlier
    print (Lr, Ur)
7. Replaces the value of outliers with median value of the features in dataset.
   median = df. loc [df ['feature'] > Ur, 'feature’].median()
   df ["feature"] = np. where(df ["feature"] > Ur, median , df ['feature'])
   median = df.loc[df['feature'] < Lr , df['feature'].median ()
   df ["feature"] = np.where(df["feature"] < Lr , median , df ['feature'])
```

The attributes after implementing operations with outliers, then go for processing to fill missing values. Null values may lead to the wrong direction for the classification process. Now, we used the imputation method to impute the missing data by the mean value of attributes instead, dropping null values in the equation (5).

$$\text{Missing\_value}=\begin{cases} mean(x), & if\ x = Null/NaN \\ x, & Otherwise \end{cases} \quad (5)$$

where x is a feature attribute value that lies in n-dimensional space. The feature scaling was picking up for normalizing the data within a range. It’s a part of the preprocessing of the dataset before applying ML techniques. The scaling of features is relevant to the independent features of the dataset. The feature scaling also is known as standardization. It also boosts the calculation speed of the algorithm. And feature selection is implemented one-way ANOVA F-test to reduce the high data dimensionality of the feature space before the classification process. It is a statistical method to select the features which highly related to the outcome variable. Then, obtained six features that highly dependent on the class label of the PIDD data. Then, we implemented an oversampling technique to integrate strongly similar data points through KNN. After applying this oversampling technique, the changes in the distribution of attributes in PIDD shown in Fig. 4. The original set of data points compared with the simple random oversampling data points and the range of training samples make larger by using the ADASYN method.

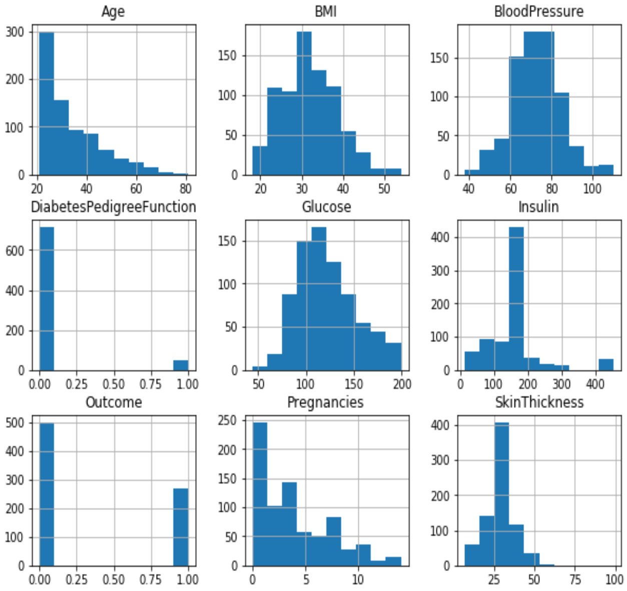

Fig. 4 Distribution of Attributes after Applying ADASYN

Finally, the machine learning-based multilayer perceptron (MLP) classifier used for predicting the classification of diabetes with the K-fold cross-validation (CV) technique. The PIDD data contains an imbalanced proportion of positive and negative data points. So, the cross-validation technique enables the capability of preserving the originality of data. The proposed model explained in Fig. 5.

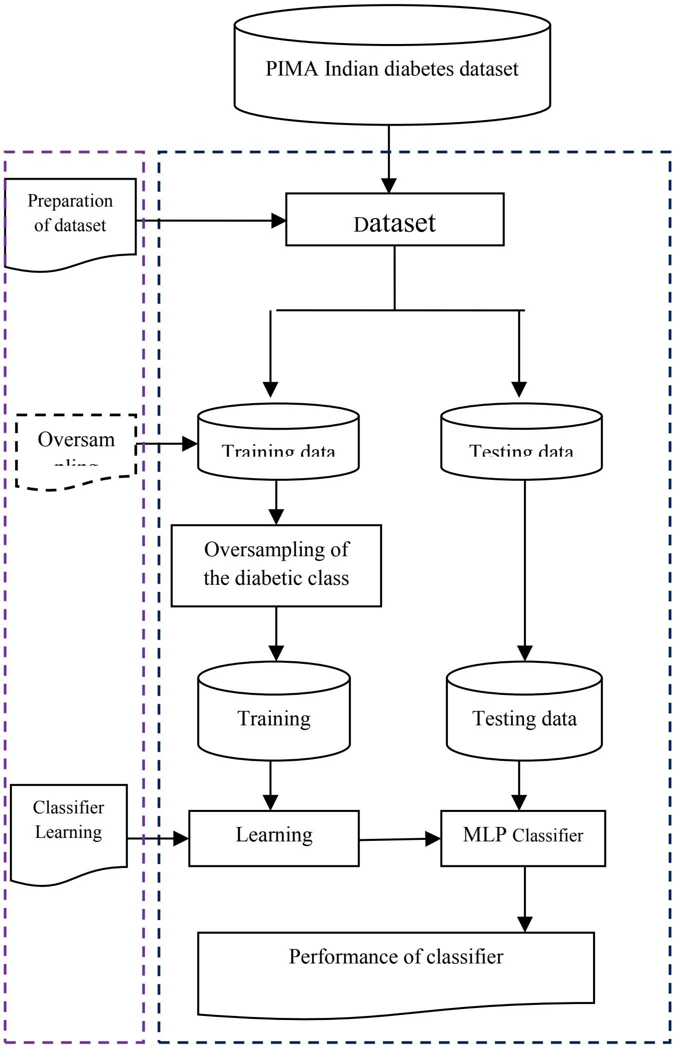

Fig. 5 Block diagram for Proposed Mechanism

V. Multilayer Perceptron

The classification task performed by the fully connected multilayer neural network. This type of neural network named as multilayer perceptron (MLP) classifier. MLP is a part of feed-forward ANN. It utilizes a supervised machine learning technique to provide training for the dataset. It comprises processing units is called neurons. In MLP, each neuron connected to another neuron with some initial weights. In this proposed model, we are using several layers as the input layer, output layer with some hidden layers. In this experiment, we used MLP classifier with five hidden layers as h1, h2, h3, h4, h5 with n1=500, n2=500, n3=500, n4=500 neurons was chosen as best architecture, while we tried with different size of neurons. We trained our MLP model on 10 epochs. In this experiment, we are learning how to optimize the hyper-parameters from this dataset. The hyper-parameters as number of neurons in each hidden layer, activation-function, learning rate, number of an epoch, the value of alpha, etc. will be used for the improvement in the performance of a classifier. Proposed mechanism mentioned in the following Algorithm 2.

Algorithm 2 Proposed algorithm
Input: Pre-Processed PIDD
Output: Classify the dataset into diabetic and non-diabetic class.

```
procedure train_test_split(feature, target, test size)
    X_train = X(total set- test set)
    X_test = X(test set)
    y_train =MapColumnWise(X_train)
end procedure
procedure Featureselection(f_classif)
    test = SelectKBest(score_func = f_classif, k = 8)
    fit = test.fit(feature, target)
    feature = fit. transform(feature)
    print(feature)
end procedure
procedure StandardScaler(sc) for  X_train, X_test
    X_train = sc.fit_transform(X_train)
    X_test = sc.fit_transform(X_test)
    Return X_train, X_test
end procedure
procedure ADASYN oversampling(X_train, y_train)
    ada = ADASYN()
    X_train, y_train = ada.fit_resample (X_train, y_train)
    return Resampled dataset shape Counter
end procedure
X_train, X_test, y_train, y_test = train_test_split (feature, target, 0.20)
Apply classifier for prediction(MLPClassifier):
mlp=MLPClassifier(hidden_layer_sizes = (500,500,500,500,500), activation='relu', solver ='adam', max_iter=10)
 mlp.fit (X_train, y_train)
pred = mlp.predict(X_test)
 end
```

VI. Evaluation Metrics

The model implemented using Python programming language. We used python programming language using a jupyter notebook environment on the version of Anaconda 3. For experiments, we selected data for training and testing purposes. The performance of the proposed model evaluated as accuracy, precision, recall, F-measures, according to the confusion matrix shown in Table 2. The confusion matrix consists True positive (TP), true negative (TN), false-positive (FP), false-negative (FN) along with different metrics. The performance of the proposed model evaluated as the results of average accuracy with the help of a k-fold cross-validation technique. Where the accuracy means the ratio of the number of correctly classified diabetes patients to the total number of diabetes datasets comprises diabetic and non-diabetic. Precision means the ratio of the number of a positive classified diabetic patient to the number of samples predicted as a patient having positive in diabetic test. Recall means the ratio of truly positive diabetes datasets to the number of the actual positive diabetic patient. F-measure means the harmonic mean of accuracy and precision which works effectively for diagnosing tests. For reporting a well-ranked prediction instead of, absolute values. It is possible with the use of the area under the curve (AUC). The results for the confusion matrix shown in Table 3.

Table 2 Confusion Matrix

Actual/ Predicted	Positive	Negative
Positive	TP	FN
Negative	FP	FP

$$Accuracy = \frac{TN + TP}{TP + TN + FP + FN} \quad (6)$$

$$Precision = \frac{TP}{TP + FP} \quad (7)$$

$$Recall = \frac{TP}{TP + FN} \quad (8)$$

$$F - measure = \frac{(2 \times Precision \times Recall)}{(Precision + Recall)} \quad (9)$$

VII. EXPERIMENT RESULTS

For all experiments, we assigned 80 percent data for training purposes and 20 percent for testing purposes. The performance of the proposed model evaluated as the average results obtained from the cross-validation technique. The accuracy of the classification task is to act as a performance indicator for the analysis of diabetes using ML techniques. Due to the presence of outliers, missing values, and imbalanced class distribution in PIDD, as a performance metric accuracy is not enough for this work. Thus, we need some points under considerations for the evaluation of the proposed algorithm is as follows-

1. Conduct classification using MLP classifier with and without detection of outliers and missing values imputation to interpret the improvement in performance after the processing of outliers and missing values.
2. Implement the classification task using MLP classifier with and without oversampling technique and feature selection to check the effectiveness of processing of class imbalance and standardization of features to analyze the likelihood of diabetes.
3. Evaluate the results against some evaluation metrics such as accuracy, recall, precision, f-measure, AUC for the all-inclusive performance of the proposed algorithm.
4. Compares the proposed algorithm with the other algorithms already used and provides better results as the performance metrics to verify that the proposed algorithm works better than those algorithms. The calculation of accuracy, precision, recall, f-measure is as below in Table 4.

Table 3 Results in Confusion Matrix

Actual/ Predicted	Diabetic	Non-diabetic
Diabetic	91	16
Non-diabetic	9	38

Table 4 Results Comparison

Method	Accuracy	Precision	Recall	F-measure	AUC score
SVM	0.73		0.65	0.52	0.51
RF	0.79	0.73	0.75	0.72	0.77
Proposed algorithm (k=5)	0.84	0.91	0.85	0.88	0.83

We analyzed the results of the proposed algorithm with other benchmark algorithms in Table 4 from the following aspects. First of all, for the MLP adopted in this paper and proposed the Algorithm with the impact of the 5-fold or 10-fold KCV strategies, the proposed algorithm gives Based on the experiment [22] by the authors, there is no processing for the outliers, missing values, and oversampling technique for the dataset. In our proposed work, we are working on these aspects and done oversampling to avoid the problem of the minority class. And also work for the feature selection method to improve the performance of the classification task.ML techniques are random forest (RF), SVM, and deep learning methods used by Yh. Amani [22] to perform a classification task for predicting diabetes on PIDD with all the performance metrics are defined above.

We compare our results of the proposed algorithm with the benchmark algorithm using the same dataset. In Fig. 6, the graphical representation shows the performance comparison between the experimental results for diabetes prediction. It shows clearly that the proposed model gives better results.

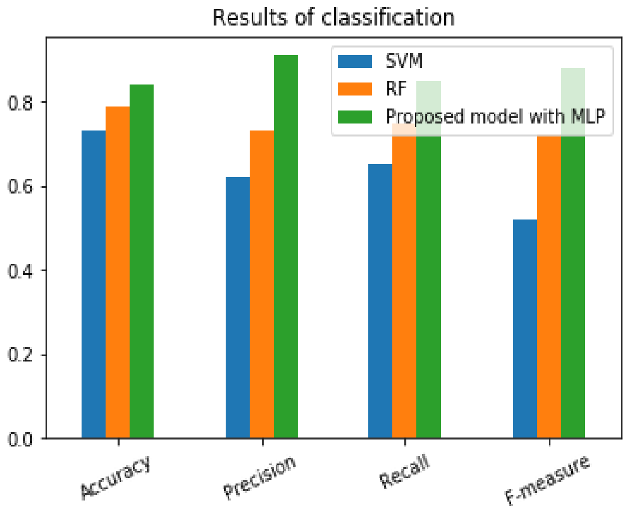

Fig. 6 Graphs Represents Comparison of Results

VIII. CONCLUSION

In this paper, we have proposed a model for classification tasks to predict diabetes. The classification method focuses on some problems regarding diabetes prediction. It considered the challenge is the presence of a missing value, outliers, and unbalances class distribution seen in the medical record. The algorithm designed for working with the data preprocessing and classification steps. The reduction of the high dimensionality of the dataset enhances the quality of dataset with decreased class imbalance where preprocessing plays a significant role in robust and precise prediction for classifying the diabetes dataset. These preprocessing steps can improve the kurtosis and skewness of the attribute distribution in PIDD data. The analysis of variance using an f-test boosts the correlation between features concerning the outcome label. The validity of the MLP classifier verified using the CV validation strategy. Due to the presence of class imbalance, other than accuracy performance metrics used as comprehensive indicators in this model. The comparison of results interprets that our proposed algorithm has outperformed the other benchmark algorithms in the experiments based on accuracy and other performance metrics of classifier; we have shown that great initiative for the prediction from the PIDD data.

REFERENCES

1. M. C. Fitzmaurice, C, L. Allen, R.M. Barter. L. Barraged, A.Z. Butte, H.N. Brenner and T.M. Fleming. Global, regional, and national cancer occurrence, mortality, vanished many years of patient, many years patient remain alive with a disability, and adjusted living for 32 cancer patient with a disability, from the 1990s: a systematized analysis for the global implication of cancer study. col. 3(4) JAMA(2017)
2. K. Sumengalli, Gitika, S.B.R., and H.Ambharkar. A classifier based method for the earliest detection of diabetic or non-diabetic. In: 2016 International Conference on Control, Instrumentation, Communication, and Computational Technologies (ICCICCT). IEEE (2016)
3. American Diabetes Association. "Diagnosis and Classification of Diabetes." Diabetes Care (May 2018)
4. C. Ambeeka and G. Dipak. A Survey on Medical Prognosis of Diabetes Using Machine Learning Methods. ©Springer Nature (2019)
5. PIMA Indian diabetes dataset: https://archive.ics.uci.edu/ml/dataset-PIMA-indian-datasets/
6. At. el .Carerra, E.V. Gonzales, R. "Automated detection of diabetic retinopathy using support vector machine, IEEE XXIV International Conference on Electronics, Electrical Engineering, and Computing (2017)
7. M. Sued, S. Lara, L. Abdurrahman,R. Almohaini, and T. Saba. Current Techniques for Diabetes Prediction: Review and Case Study. Applied. Science. (2019)
8. Md. F. Faisal, H. Ashaduzaman, Sharker, Iqbal. Performance Analysis of Machine Learning Techniques to Predict Diabetes. IEEE International Conference on Electrical, Computer and Communication Engineering (2019).
9. P.Shonar K.MaliniJaya, Diabetes Prediction Using Different ML Approaches.3rd International Conference on Computing Methodologies and Communication (ICCMC), (2019)
10. D. Vigneshvaari, N. K. Komal, R.Ganesh, V, A. Gagan, Vikas, S. R. Machine Learning Tree Classifiers in Predicting Diabetes. 5th International Confer- ence on Advanced Computing Communication Systems(2019)
11. Jurafsky, S. Denials H. Martin, A. James, Logistic regression: Speech and Language Processing(2019)
12. S.S.Dileep, S.Deepti. Prediction of Diabetes using Classification Algorithms. International Conference on Computational Intelligence and Data Science (IC- CIDS 2018)
13. M. Maniruzzaman, M. J. Rahman, M. Al- MehediHasan, H. S. Seri, M. M. Abedin, A. El-Baz, and J. S. Seri. Accurate diabetes risk stratification using machine learning: Role of missing value and outliers. J. Med. Syst., vol. 42, no. 5, p. 92,(May 2018).
14. H. Kaur and V. Kumari. Predictive modeling and analytics for diabetes was using a machine learning approach. Appl. Compute. Information. (Dec. 2018)
15. K. Pradhnya, B.Rahul: Analysis of Classifiers for Prediction of Type II Diabetes.In Proceedings of International Conference on Computing Control and Automation (ICCUBEA), IEEE(2018)
16. K.Rolikathi, R, RajaKumari, P,G. Prashant. Diagno- sis of Diabetes Using Cascade Correlation andANN. Tenth International Conference on Advanced Com- puting (ICoAC). (2018)
17. G. Sapna, R. Vinnie Kumar, and K. P. Soman. Dia- betes detection using deep learning algorithms. ICT Express, vol. 4, no. 4, pp. 243–246 (2018)
18. Sneha, N., and T. Gangil. Analysis of diabetes mellitus for early prediction using optimal features selection. Journal of Big Data 6 (1): 13, (2019)
19. Scikit- learns and related packages: https://scikitlearn.org/ stable/ modules
20. Yoho. Amani, Jm. Akhtar, Yesiltepe Misrad, R.Jawad. "A decision support system for diabetes prediction using machine learning techniques and deep learning technique", pp. 1-6, 2019.