Improving the Diabetes Diagnosis Prediction Rate Using Data Preprocessing, Data Augmentation and Recursive Feature Elimination Method

E. Sabitha, M. Durgadevi
Dept. of Computer Science and Engineering, College of Engineering and Technology
SRM Institute of Science and Technology Vadapalani, Campus No.1, Jawaharlal Nehru Road, Vadapalani TN, India

***Abstract*—Hyperglycemia is a symptom of diabetes mellitus, a metabolic condition brought on by the body's inability to produce enough insulin and respond to it. Diabetes can damage body organs if it is not adequately managed or detected in a timely manner. Many years of research into diabetes diagnosis has led to a suitable method for diabetes prediction. However, there is still scope for improvement regarding precision. The paper's primary objective is to emphasize the value of data preprocessing, feature selection, and data augmentation in disease prediction. Techniques for data preprocessing, feature selection, and data augmentation can assist classification algorithms function more effectively in the diagnosis and prediction of diabetes. A proposed method is employed for diabetes diagnosis and prediction using the PIMA Indian dataset. A systematic framework for conducting a comparison analysis based on the effectiveness of a three-category categorization model is provided in this study. The first category compares the model's performance with and without data preprocessing. The second category compares the performance of five alternative algorithms employing the Recursive Feature Elimination (RFE) feature selection method. Data augmentation is the third category; data augmentation is done with SMOTE Oversampling, and comparisons are made with and without SMOTE Oversampling. On the PIMA Indian Diabetes dataset, studies showed that data preprocessing, RFE with Random Forest Regression feature selection, and SMOTE Oversampling augmentation can produce accuracy scores of 81.25% with RF, 81.16 with DT, and 82.5% with SVC. From Six Classifiers LR, RF, DT, SVC, GNB and KNN, it is observed that RF, DT, and SVC performed better in accuracy level. The comparative study enables us to comprehend the value of data preprocessing, feature selection, and data augmentation in the disease prediction process as well as how they affect performance.**

Keywords*—Artificial Intelligence (AI); Machine Learning (ML); Deep Learning(DL); Neural Network; Diabetes Mellitus; Recursive Feature Elimination (RFE); Synthetic Minority Over-sampling Technique (SMOTE)

I. Introduction

A metabolic disorder known as diabetes mellitus is characterized by hyperglycemia brought on by the body's inability to create and utilize insulin.[1]. There are three forms of diabetes types. The human body cannot generate enough insulin when it has type I. The body is unable to produce or use insulin effectively in type II. During pregnancy, gestational diabetes can develop [2]. Both Type I and Type II diabetes are getting more and more prevalent worldwide, with Type II diabetes being at epidemic levels. According to medical study, diabetes has been linked to the long-term degradation of vital organs. More concerning is its impact on pregnancies: diabetes affects roughly 7% of pregnancies each year, posing a dual life-threatening risk Over half of the world's population is expected to have diabetes by 2045 due to the disease's rising prevalence. The WHO predicted that 463 million people will be diabetic worldwide by 2020, and these are just the cases that have been identified. In United States almost one in every 10 individuals is diabetic. Diabetes research is therefore essential, including studies of diabetes prediction and its effects on health [3].

Diabetes can be diagnosed by either an oral glucose tolerance test result or a fasting plasma glucose level. On the other hand, diabetes can be identified by Glycemic threshold levels .This is due to the fact that different ethnic groups have varied risk levels. Multiple blood sugar tests are taken both before and after a meal. By observing a relevant decision at a time, practitioners are faced with the difficult task for diagnosing diabetes. The diagnostic process, on the other hand, can be made more computationally simple [4].The fields of technology and medicine have been profoundly affected by big data and data analytics approaches. Rather than depending on conventional methodologies, which are usually unable to handle massive data, cutting-edge technologies like ML, DL and cloud computing must be employed to fully utilise the data and automate computation processes in medical research. This paper provides a customised hybrid model of artificial neural networks (ANN) and genetic algorithms as a framework for accurately forecasting the onset of diabetes, replete with regularisation and prediction techniques created for diabetes prediction [5].

Numerous computational projects have been started recently, many of which are focused on the use of ML and DL algorithms in diabetes research with the goal of assisting physicians in making rapid and accurate diagnosis decisions. With the ongoing development of diabetes testing equipment, individuals can now take part in individualised examinations of their diabetes status for better lifestyle modifications. In comparison to existing methods, a dependable accuracy rate is categorized in recent studies. A higher accuracy rate in diabetes prediction is essential, as early diagnosis of diabetes mellitus is required. The researchers are presenting a range of DL and ML methods for diabetes forecasting. Despite a large amount of research on diabetic prediction, the accuracy still

needs to be improved. This is necessary since diabetes poses major health risks if it is not effectively treated or diagnosed in a timely manner. In this paper a comparative evaluation is done based on feature selection approaches and data augmentation techniques that increase prediction performance in this research. The main contribution of the paper is summarised as follows:

1) The significance of data pre-processing is demonstrated by comparing the outcomes of the proposed model with and without data pre-processing.

2) To emphasise the significance of feature selection in disease prediction, which improves model performance and boosts predictive power.

3) To overcome the issue of a small dataset, data augmentation is employed to enhance the dataset's size. Deep learning and machine learning typically require a large quantity of data to train the networks.

II. Related Work

The views of data pre-processing, data augmentation, and classification are the foundation of the current body of work. However, in this paper the review is limited to the recent publications. Diabetic research has recently begun to improve based on the performance accuracy. This article can be used by readers to learn about the past and present effectiveness of algorithms in diabetes research [6]. Table I illustrate the review on recent papers that work on the diabetic prediction. The review is based on the recent trends and papers that are suitable for the disease prediction. In diabetes research, the NN-based approaches have constant to increase accuracy. The problems of data standardisation, imbalance, and feature augmentation are addressed in this Min-Max Normalisation and a Variational Autoencoder [7]. MLP was then utilised for classification, with an accuracy rate of 92.31 percent. In [8], the accuracy of their Artificial Backpropagation Scaled Conjugate Gradient Neural Network (ABP-SCGNN), which was previously reported to attain 93 percent accuracy without data pre-processing, has significantly improved. The work of [9] demonstrates another impressive result with NN-based models. They looked at iterative imputers, k-nearest neighbour (K-NN), and median value imputation. In order to acquire an F1-score of 98 percent, MLP was then employed for classification. For feature selection and missing value imputation in [10], Pearson correlation and median value imputation were used. The authors used interquartile ranges to further normalise the data and eliminate outliers. DNN-based classification model, which contained a number of hidden layers, was 88.6% accurate. A deep neural network (DNN) model's accuracy was estimated to be 98.07 percent in [11]. Even though the authors claim to have used data cleansing, the process is not described in the article. In [12] used the median value for missing value imputation and principal component analysis (PCA) for feature selection. MLP was then used to carry out the classification procedure, and it had a 75.7 percent accuracy rate. For feature selection and missing value imputation, PCA and minimum redundancy, maximum relevance (mRMR) was also used in [13]. Using an MLP, they were able to get a classification accuracy of 73.90%. Many different methods were employed [14]; implemented many assessment techniques, including Nave Bayesian, Random Forest (RF), KNN, and K-fold Cross-Validation. The technique has a 64.47 percent accuracy, according to K-fold Cross Validation. Nave Bayes, function-based multilayer perceptrons, and RF based on decision trees were all employed in [15]. The feature extraction method was utilised to extract reliable and illuminating properties from the dataset using the correlation method. According to the author, the Nave Bayes method outperformed the random forest and multilayer perceptron algorithms.

In [16] tested various machine learning methods for predicting early diabetes on the PID dataset. Using 20-fold cross-validation and a 70-30 train-test split, tree-based RF scored 75.65 percent, Nave Bayes (NB) 71.74 percent, and KNN 65.19 percent. A decision tree and the gradient boosting method were used by [17] for prediction. The technique has a classification accuracy of 90% and computes a correlation value to determine differences between a diabetic patient and healthy person. With 10-fold-cross validation and an enhanced K-Means cluster method, [18] obtained 95.42 percent accuracy. In [19] used 10-fold cross validation with machine learning techniques on patients who had a history of non-diabetes and a cardiac problem. In [20], which used machine learning as a prediction model for type 2 diabetes mellitus early prediction, Glmnet, RF, XGBoost, and Light all shown improved clinical prediction. It is suitable for one dataset but inappropriate for another. A more advanced DNN-based diabetes risk prediction model that not only predicts but also identifies who will develop the ailment in the future was proposed in [21]. Before training on several classification models, such as NB, LR, RF, AB, GBM, and extreme gradient boosting, the mean of each column of data was pre-processed in [22] to remove missing values. With a precision of 77.54 percent, the XGBoost model was the most precise. The efficiency of the classification models SVM, K-NN, NB, Gradient boosting (GB), and RF were contrasted in [23]. With an accuracy of 98.48 percent, the RF prevailed. In [24] employed Pearson correlation for feature selection and mean value imputation for missing value. The authors assessed the performance of various classification models, including extreme boosting (XB), AB, RF, DT, and K-NN, using a K-fold cross-validation environment and the grid search strategy for hyperparameter tuning. With an accuracy percentage of 94.6 percent, the XB won. Linear SVM, Radial Basis function SVM, DT, and K-NN were employed in a stacked ensemble to achieve a classification accuracy of 83.8%.

TABLE I. OVERVIEW OF THE LITERATURE REVIEW

Authors	Feature selection (FS)	Classification	Comments
Benavides, C., et al .[7]	FS: none specified;removed missing values;	MulltiLayer Perceptron	MLP achieved the best accuracy, 92.31%
Alkhamees, B. F et al.[8]	FS: none specified; MVI: none specified	ANN trained with ABS conjugate gradient neural network (ABP-CGNN)	Achieved 93% accuracy
Ahmad, M., et al.[9]	Median value, K-NN, and iterative imputer were used for missing value imputation	ANN	ANN achieved 98% accuracy
Foo, S. Y et al.[10]	FS: Pearson correlation MVI: Median value for missing values imputation.	DNN run with different hidden layers	Achieved 86.26% accuracy with 2 hidden layers.
Naz, H., & Ahuja, S. [11]	Method not stated	MLP and DL with 2 hidden layers	DL achieved best accuracy of 98.07%
Iqbal, M. A., [12]	FS: PCA; MVI: Median value	MLP	Achieved 75.7% accuracy
Qu, K., et al. [13]	FS: PCA; MVI: redundancy and minimum relevance	MLP	Achieved 73.90% accuracy
Halgamuge, M. N., et al.[14]	none specified	NB,RF,KNN,K-fold crosss validation	Using K-fold CrossValidation, the method achieved 64.47% accuracy.
Singh, D. A. A. G., et.al.[15]	Correlation method	decision tree-based RF, function-based multilayer perceptron ,Naïve Bayes	Naïve Bayes algorithm achieved better results
Awais, M., et.al.[16]	none specified	RF,NB,KNN with 20 -fold cross-validation	RF achieved 75.65%
Selvan, K. A., et.al.[17]	none specified	DT,GB	Achieved 90% accuracy
Yang, S., et.al.[18]	none specified	K-Means cluster algorithm with 10 fold-cross validation.	Acheived 95.42% accuracy
Gnanadass [22]	none specified	NB, linear regression (LR), RF, AB, gradient boosting machine (GBM), and extreme gradient boosting (XGBoost).	XGBoost achieved 77.54%
Mounika, B., et al.[23]	none specified	SVM, K-NN, NB, GB, RF, LR	RF achieved best accuracy of 98.48%
Hasan et al [24].	FS: correlation; MVI: mean value	XB, AB, RF, DT, K-NN	XB achieved best accuracy of 94.6%

III. MATERIALS AND METHODOLOGY

A. Dataset

The PIMA Indian Diabetes database was used for this study. The main objective of the dataset is to establish a patient's diagnostic diabetes status. The dataset contains one outcome variable and a number of medical predictor variables. Predictor variables for diabetes include age, number of pregnancies, BMI, BP, glucose, Skin thickness, Insulin, and Diabetes pedigree function. Particularly, all of the patients are females in PIMA who are at least 21 years old. The selection of these examples from a broader database was subject to several of limitations. Our proposed research compares data pre-processing, feature selection, and data augmentation techniques. The study aims to overcome flaws in early diabetes mellitus diagnosis that impair accuracy. The disadvantages are as follows: 1. A large number of missing values lead to erroneous predictions. 2. Imbalanced data has an impact on the model's performance. The suggested framework illustrates each stage of prediction work, including data pre-processing, Feature selection techniques incorporated with the Recursive Feature Elimination approach, and Smote data augmentation with and without Smote data. The study compares model accuracy by applying the augmentation strategy to improve the dataset. The study compares not only on the basis of augmentation, but also on the basis of feature selection strategies. Fig. 1 shows a diagrammatic representation of the planned work.

IV. PROPOSED WORK

In this paper, a systematic framework is proposed for the diabetic prediction with different classifiers. The importance of the data preprocessing is elaborated by presenting the results obtained. The main contribution of the proposed work is to show the importance of data cleaning by using the data preprocessing strategies then by selecting the important attributes that highly correlate with diabetic prediction. Most important part is balancing the dataset by SMOTE data augmentation, the proposed work focus on the three part that improves the prediction accuracy. The six classifier algorithm are then used for classification. In the proposed system, inconsistent data is replaced, then suitable features are selected by using the RFE with the Random Forest Regression and finally the selected are augmented by SMOTE

Oversampling technique to improvise the imbalance dataset problem.

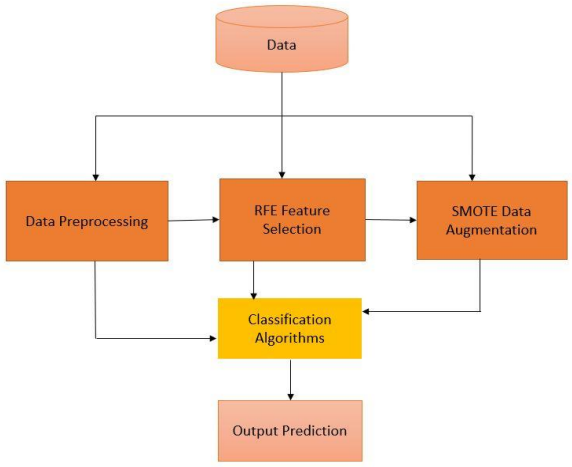

Fig. 1. Proposed Work.

V. Data Pre-Processing

The idea of data pre-processing refers to the conversion of unclean data into a clean data set. The dataset is pre-processed to look for missing values, noisy data, and other irregularities before applying the algorithm. These data are crucial for decision-making, thus accurate and effective estimate techniques are required. For this analysis, PIMA Indians Diabetes database is taken. The dataset has more missing values than null values. In the medical field, the problem of a database with missing values is very widespread. The Table II displays the number of zeros in each attribute, while Fig. 2 and Fig. 3 illustrate the proportion of missing data and the impossible value assigned in the pregnant feature, both of which reduce the model's performance. Using data pre-processing techniques, the study focuses on cleaning up the data by improving the values assigned to each feature. In this paper no specific strategies are followed to clean up the data in the suggested work, instead a few simple and easy ways to pre-processing is done to clean up and improve the quality of the data. The following are the methods that will be described.

TABLE II. Missing Value In Dataset

Features	Total	Percent
Insulin	374	48.697917
Skin Thickness	227	29.557292
Blood Pressure	35	4.557292
BMI	11	1.432292
Glucose	20	0.651042
Pregnancies	0	0.00
Diabetes Pedigree Function	0	0.00
Age	0	0.00
Outcome	0	0.00

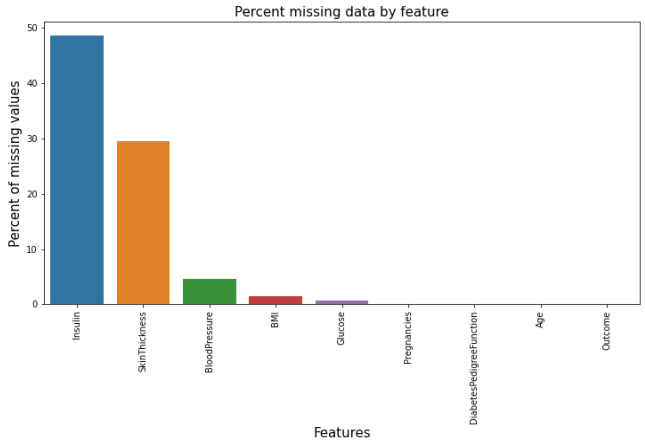

Fig. 2. Percent of Missing Data in Features.

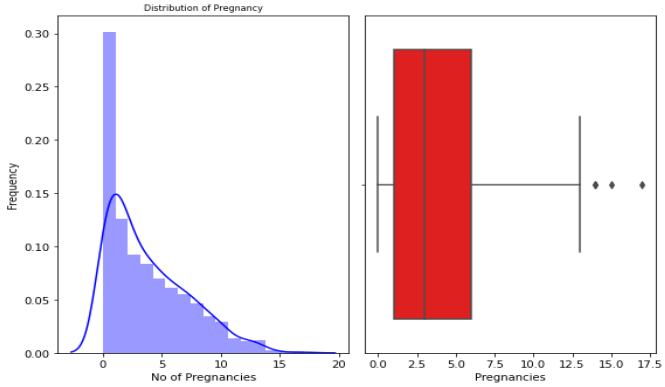

Fig. 3. Outliers of Pregnancies Feature.

A. Treatment of Missing Values

If the behaviour and linkages with other variables are not adequately analysed, missing data in a data collection could reduce a model's power or fit or result in a biased model. The classification or prediction that results could be inaccurate. When the dataset's above mentioned attributes were evaluated, it was found that several of them had zero values and that the pregnancy variable had a maximum value of 17, which seemed to be impossibly high. There is a range for a typical healthy human being that is not zero, suggesting a missing value, hence these 0 column values are illogical. To make counting the missing integers simpler, we'll start by swapping out these zeros for NaN. Later, we'll swap them out for the proper values. There are a number different ways to handle missing values. The choices are displayed below. Simply deleting all instances where the variable being considered contains missing values is the simplest method to handle with missing values. However, this approach can mean losing actually valued data about patients. The calculation of mean is the second approach to complete all gaps in the data. In this approach the missing values are replaced by with the average value calculated in the same attribute. This method can reduce the loss of data instead of removing the missing values that reduces the quality of the dataset. In this method all the missing values are replaced by zero. The process of replacing by zero is simply by replacing any missing values with zero. Since the data in this study have been converted to values

between zero and one, substituting zero for missing values has the same result as substituting the attribute's lowest value. However this method leads to poor classification by missing data which are inaccurately appraised if they are necessary for clinical management. The K-nearest neighbour method is the fourth method which is used to replace the missing values in the dataset .The missing values are replaced with the value of nearest-neighbor column. The nearest-neighbor column is considered to be the closest column in Euclidean distance. The next closest column is used if the relevant value from the nearest-neighbor column also contains a missing value.

The Table I clearly describes the PIMA dataset. The features like Glucose, Blood Pressure, Skin thickness, Insulin, and BMI are with 0 values. In this study, we use second approach i.e. all missing values of an attribute are replaced by the mean by calculating the average of all accessible values of the same attribute. When all the zero values are replaced with mean value, the dataset was further split into training and testing data. The dataset as a whole is made up of 80% training data and 20% test data. The model performance is evaluated by the model accuracy, which is determined via machine learning algorithms. On two levels—one where the zero values were replaced with the mean and another where they weren't—we compared the model's performance. By contrasting the two, we can see how useful data pre-processing is in improving the dataset's suitability for subsequent operations. The comparison can be seen in upcoming session.

B. Without Data Pre-Processing

The PIMA dataset is directly used in the machine learning algorithm to assess the prediction's accuracy without any data pre-processing. Some of the techniques used are Gaussian Nave Bayes (GNB), KNN, DT, Support Vector Classification (SVC), LR, and RF. The dataset is used in the classification approach to determine the degree of accuracy in disease prediction that may be made without any prior processing. Table III and Fig. 4, clearly represent the performance of the classification algorithm based on the accuracy. The accuracy for LR was 77.5 percent with 0.06 training time, for RF it was 76.5 percent with 0.77 training time, for DT it was 67.5 percent with 0 training time, for SVC it was 81.2 percent with 0.03 training time, and for GNB and KNN it was 75% with 0.02 training time.

TABLE III. ACCURACY OBTAINED WITHOUT DATA PRE-PROCESSING

S.no	Classifier	Validation accuracy	Training time
0	Logistic Regression	0.775	0.06
1	Random Forest	0.7625	0.77
2	Decision Tree	0.675	0
3	SVC	0.8125	0.03
4	GaussianNB	0.75	0.02
5	KNeighbors	0.75	0.02

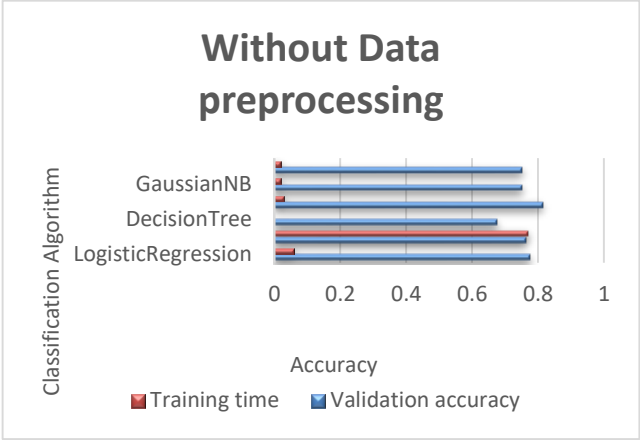

Fig. 4. Accuracy Without Data Pre-processing

C. With Data Pre-processing

1) Dropout missing values: The simple method to deal with missing values, is to simply delete any instances in which the variable being studied contains missing values. The loss of potentially relevant data regarding patients whose values are missing, however, could be a consequence of this strategy. The dataset is divided in an 80:20 ratio between training and testing data, after all missing values have been removed, and the features are chosen using RFE with Random Forest regression. An imbalance dataset generally speaking is the PIMA dataset. We can see that 268 people have diabetes and 500 people do not when we analyse the dataset by outcome.

The performance of the classification algorithm worsens when the data for training and testing are split due to the unequal size of the training and testing sets. Using classification techniques to determine correctness, we augment the dataset with additional data to address the imbalance issues. Table IV and Fig. 5 represent the result obtained. The accuracy scores are LR 69.04 percent with 0.05 training time, RF 85.7 percent with 0.62 training time, DT 73.8 percent with 0.0 training time, SVC 76.1 percent with 0.01 training time, GNB 73.8 percent with 0.02 training time, and KNN 71.42 percent with 0.01. When we focus on accuracy, RF has achieved the maximum score of 85.7 percent, but the training duration is 0.62 minutes. When it comes to training time, DT, SVC, GNB, and KNN take less time, but their accuracy is worse than RF.

TABLE IV. ACCURACY OBTAINED WITH DROP OUT MISSING VALUES

S.no	Classifer	Validation accuracy	Training time
0	Logistic Regression	0.690476	0.05
1	Random Forest	0.857143	0.62
2	Decision Tree	0.738095	0
3	SVC	0.761905	0.01
4	GaussianNB	0.738095	0.02
5	KNeighbors	0.714286	0.01

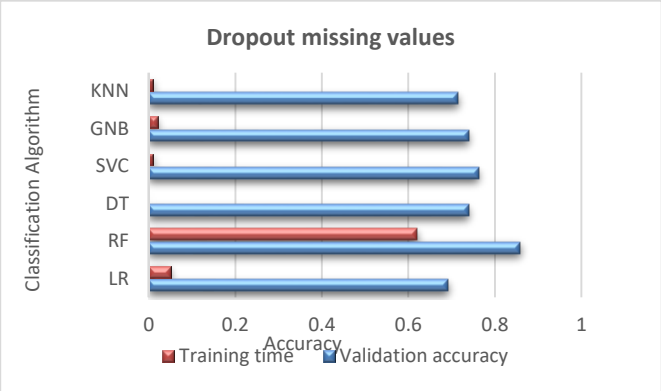

Fig. 5. Accuracy with Dropout Missing Values.

2) Replacing missing value with mean: There are no null values in the PIMA dataset, however there are more missing values, such as zero values in many characteristics, as previously indicated. As a result, the model's performance may be affected. To solve this problem, one option is to remove the zero values, but this reduces the algorithm's performance. Instead, in this study, we can replace the zero by calculating the attribute mean values and replacing the zero. One way to pre-process a dataset without reducing its size is by using this technique. Following pre-processing, the dataset is divided into train and test groups in an 80:20 ratio. The next stage is to assess the accuracy of the diabetic prediction using classification algorithms.

TABLE V. ACCURACY OBTAINED WITH MEAN VALUE

S.no	Classifier	Validation accuracy	Training time
0	Logistic Regression	0.725	0.04
1	Random Forest	0.7625	0.52
2	Decision Tree	0.7	0
3	SVC	0.7625	0.02
4	GaussianNB	0.7	0.01
5	KNeighbors	0.7625	0.01

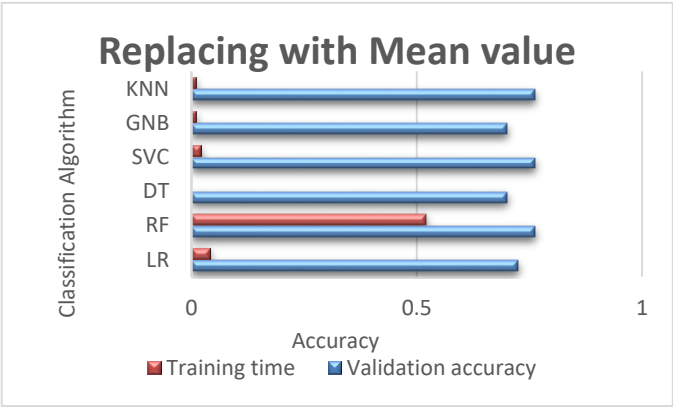

Fig. 6. Accuracy with Replacing Mean Value.

The accuracy of classification algorithms when the inconsistent value is replaced by Mean value is explained in Table V and graphically presented in Fig. 6. When the dataset is pre-processed, we can witness a gradual improvement in accuracy in Tables IV and V. The dataset was pre-processed in this study by removing zero values and replacing them with the mean of the characteristics. There isn't much of a change, however pre-processing the dataset can help us gain a little more precision. Understanding the significance of data pre-processing in any research effort is made easier by this approach. Data pre-processing is essential for evaluating the model's performance or determining the algorithm's efficiency.

VI. FEATURE SELECTION

Finding the most useful set of features for creating efficient models of the phenomenon being studied is the goal of feature selection. Feature selection strategies are divided into two categories: i) supervised techniques and ii) unsupervised procedures. The efficiency of supervised models is increased by using labelled data in supervised procedures to find pertinent characteristics. Unlabelled data is utilised in unsupervised approaches. In terms of taxonomic classification, these methods fall under the headings of A) Filter methods, B) Wrapper methods, C) Embedded methods, and D) Hybrid methods.

A. Recursive Feature Elimination

To choose the features in the study, a Wrapper method based Recursive feature elimination (RFE) strategy is applied. RFE is a greedy optimization strategy that selects features by considering a reduced set of features iteratively. A variety of deep learning algorithms are provided and employed in the method's core to choose features. On the other hand, filter-based feature selections rank each feature according to its importance and choose the ones with the highest or lowest scores. The given algorithms, such as random forest, decision trees, and SVM, are used to score features, or a more general technique that is independent of the whole model is used. The importance of the features used in training the estimator is decided using the feature importance attribute. Until we get the required number of features, the least important feature is deleted from the existing collection of features.

Procedure

Step 1: Fit the RFE method to the model

Step 2: The feature importance attribute is used to rank features.

Step 3: Once the necessary number of features is collected, the least significant feature is deleted and the procedure is repeated.

RFE with five different algorithms

In this study feature selection done five different ways.

1) Manual feature selection
2) RFE with Logistic regression
3) RFE with Random Forest regression
4) RFE with Decision Tree regression

5) RFE with Decision Tree Classifier
6) RFE with 5-Cross validation

B. Manual Feature Selection

The features for the prediction are manually picked in manual feature selection. The outcome is one of nine attributes in PIMA, eight of which are independent. We chose six attributes out of the eight for this research. The characteristics were chosen after reviewing a large number of research articles that were more accurate in diabetic prediction. We choose six attributes manually like Pregnancy, Glucose, Blood Pressure, Insulin, BMI, and Age. The dataset is then split in half, 80:20, into train and test sets. The classification algorithms such as GNB, KNN, DT, SVM, and RF are utilized, to analysis the performance quantified in terms of accuracy. The dataset, methodology, and accuracy achieved are all evaluated in the Table VI (also see Fig. 7).

C. Recursive Feature Elimination with different Methods

In this research, RFE was used in conjunction with several methods such as logistic regression. The key characteristics from the dataset are selected using Random Forest regression, Decision tree regression, Decision tree classifier, and RFE with five-fold cross validation as an estimator. Pregnancies, BMI, DPF, and Glucose are the most common features selected by each feature selection technique. Pregnancies, BMI, DPF, and Glucose are four of the eight variables that are thought to be directly connected to diabetic prediction. The dataset is then divided into a train set and a test set in an 80:20 ratio based on the selected attributes. Classification algorithms like GNB, KNN, DT, SVM, and RF are used for prediction, and the performance of the algorithm is evaluated in terms of accuracy. The dataset, feature selection procedure, selected features, Classification Algorithm, and Accuracy are all detailed in Table VII

TABLE VI. MANUAL FEATURE SELECTION

Dataset	Classification Algorithm	Accuracy
PIMA	GNB	77.27%
	KNN	75.97%
	DT	70.10%
	SVM	79.87%
	RF	81.81%

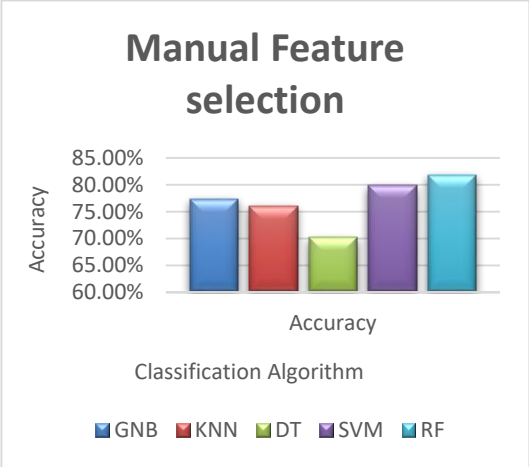

Fig. 7. Accuracy with Manual Feature Selection.

RFE with Logistic Regression selected Pregnancies, BMI, and DPF features after observing several feature selection methods. The features are then divided into two groups: training and testing. For classification and accuracy evaluation, algorithms such as GNB, KNN, DT, SVM, and RF are utilized. GNB obtained 79.87 percent accuracy using the algorithms. RFE with Random Forest Regression, using Glucose, BMI, and DPF as selected features, has a greater accuracy of 81.16 percent when compared to DT. The same features were chosen from RFE with Decision Tree Regression and RFE with Decision Tree Classifier. KNN achieved 75.32 percent accuracy by using both selection methods. Pregnancies, Glucose, and BMI were chosen as characteristics for RFE with five cross validations. The KNN algorithm achieved 79.2 percent accuracy based on the features. When compared to other algorithms, RFE with Random Forest Regression utilizing DT has obtained greater accuracy of up to 81 percent, according to the detailed analysis using the RFE feature selection method and algorithm.

VII. DATA AUGMENTATION

Data augmentation is a series of techniques for generating extra data points from existing data in order to fictionally increase the amount of data accessible. Simple data modifications or the use of deep learning models to generate more data are instances of this. Applications for machine learning are quickly increasing and diversifying, especially in the deep learning space. Approaches for data augmentation may be effective in the struggle against the drawbacks of artificial intelligence. One step in building a data model is cleaning the data, which is necessary for high accuracy models. The model won't be able to produce reliable predictions for inputs from the real world, though, if data cleansing limits representability. In order to increase the reliability of machine learning models, data augmentation techniques can be employed to replicate variations that the models would encounter in the actual world.

A. SMOTE Oversampling

In many disciplines, unbalanced data has been a problem, causing most approaches to produce erroneous forecasts that strongly favour the dominant class. To decrease the harmful impact of unbalanced data, we can optimise the process using a variety of techniques: Certain techniques, like as oversampling, under sampling, or both, are employed to correct the unbalanced data set in order to generate a balanced distribution. A statistical method for equally expanding the number of cases in a dataset is called SMOTE (Synthetic Minority Oversampling Technique). Based on current minority conditions, the component creates new instances. The overfitting issue brought on by random oversampling is helped by the SMOTE algorithm. The working approach begins by setting up the total number of oversampling observations N. A binary class distribution of 1:1 is typically used to select it. This could be minimised, though, depending on the circumstances. After that, a positive class instance is randomly chosen and the loop starts. Then, the KNNs for that instance are obtained. In order to interpolate new synthetic

instances in the end, N of these K instances are chosen (see Fig. 8).

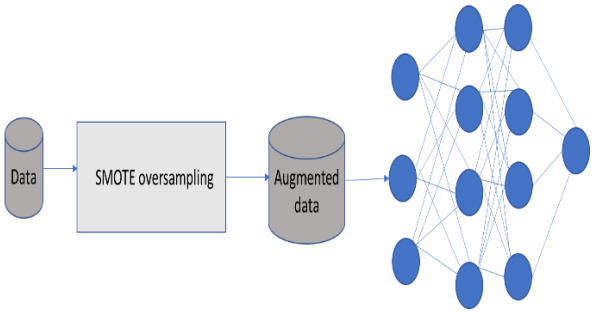

Fig. 8. SMOTE Oversampling.

Algorithm

Step 1: To interpolate new synthetic instances in the end, N of these K instances are chosen. x∈A.

Step 2: The uneven proportion determines the sampling rate N. N samples (x1,x2..xn) are chosen at random from x∈A Set A1 is composed of the k-nearest neighbours.

Step 3: For each xk, a new example is produced using A1 (k = 1, 2, 3,... N): x^'=x+rand(0,1)*|x- xk | in which rand(0,1) represents the random numbers between 0 and 1.

B. SMOTE Oversampling with RFE with Random Forest Regression

In the PIMA dataset, the result is in an unbalanced state. When examining the outcome, 1 counts to 268 and 0 counts to 500, resulting in false in excess of true. The model is trained with a higher percentage of false values than true values. SMOTE oversampling is used to reduce the data's complexity and balance it. The features selected through the RFE with Random Forest regression feature selection technique are subjected to oversampling. The Random Forest algorithm includes a feature importance calculation that can be done in two ways. The Gini coefficient is calculated using the Random Forest structure. Decision Tree algorithm with internal nodes and leaves make up each decision tree that makes up a Random Forest. The internal node uses the chosen characteristic to determine how to divide the data set into two sets with similar replies. For classification tasks, criteria like gini impurity or information gain, as well as variance reduction for regression, are used to select the internal node properties. The importance of a feature is determined by the average of all trees in the forest. There is also Mean Decrease. Accuracy is a method for calculating the importance of features on permuted out of bag samples based on the accuracy's mean reduction. The scikit-learn package does not include this function. The selected features from the RFE using Random Forest Regression are Glucose, BMI, and DiabetesPedigreeFunction. The features are then enhanced based on their results, and performance is measured using machine learning algorithms such as Logistic Regression (LR), RandomForest(RF), DecisionTree(DT), SVC, GaussianNB, and KNeighbor's for further classification. The accuracy attained with and without data augmentation is shown in the table VIII and IX.

TABLE VII. RFE WITH DIFFERENT FEATURE SELECTION METHOD

Dataset	Feature selection method	Selected Features	Classification Algorithm	Accuracy
PIMA	RFE with Logistic Regression	Pregnancies BMI DPF	GaussianNB(GNB) Knearestneighbor(KNN) Decision Tree(DT) SupportVectorMachine(SVM) RandomForest(RF)	GNB: 79.87 % KNN: 72.72 % DT: 62.98 % SVM: 72.07 % RF: 71.42%
	RFE with Random Forest Regression	Glucose BMI DPF	GaussianNB(GNB) Knearestneighbor(KNN) Decision Tree(DT) SupportVectorMachine(SVM) RandomForest(RF)	GNB : 74.02 % KNN : 72.77 % DT : 81.16 % SVM : 75.32 % RF : 75.97%
	RFE with Decision tree Regression	Glucose BMI DPF	GaussianNB(GNB) Knearestneighbor(KNN) Decision Tree(DT) SupportVectorMachine(SVM) RandomForest(RF)	GNB : 74.02 % KNN : 75.32 % DT : 68.27 % SVM :74.67% RF : 72.72%
	RFE with Decision tree Classifier	Glucose BMI DPF	GaussianNB(GNB) Knearestneighbor(KNN) Decision Tree(DT) SupportVectorMachine(SVM) RandomForest(RF)	GNB : 74.02 % KNN : 75.32 % DT : 68.27 % SVM : 74.67% RF : 72.72%
	RFE with five cross validations	Pregnancies Glucose BMI	GaussianNB(GNB) Knearestneighbor(KNN) Decision Tree(DT) SupportVectorMachine(SVM) RandomForest(RF)	GNB : 74.02 % KNN : 79.2 % DT : 70.7 % SVM : 75.9% RF : 72.72%

TABLE VIII. WITH SMOTE AUGMENTATION

S.no	Classifier	Validation accuracy	Training time
0	Logistic Regression	0.80	0.03
1	Random Forest	0.7875	0.72
2	Decision Tree	0.725	0
3	SVC	0.825	0.03
4	GaussianNB	0.775	0.02
5	KNeighbors	0.7125	0.02

TABLE IX. WITHOUT SMOTE AUGMENTATION

S.no	Classifier	Validation accuracy	Training time
0	Logistic Regression	0.69	0.05
1	Random Forest	0.70	0.59
2	Decision Tree	0.72	0
3	SVC	0.7096	0.02
4	GaussianNB	0.6774	0.02
5	KNeighbors	0.6935	0.02

We can compare the accuracy gained with and without augmentation using the two tables above. The accuracy and training duration are shown in Table VIII with moderate improvement. The accuracy and training time are displayed in Table IX without any augmentation. When we compare the two tables, it's evident that by supplementing the data, we can improve accuracy. When compared to various machine learning algorithms, SVC with smote augmentation had the highest accuracy of 82.5 percent with a training time of 0.03. Whereas Logistic regression scored 80 percent with a training time of 0.03, Random Forest scored 78.75 percent with a training time of 0.72, Decision Tree scored 72.5 percent with a training time of 0, GaussianNB scored 77.5 percent with a training time of 0.02, and Kneighbors scored 71.25% with a training time of 0.02.

VIII. RESULTS

The work is primarily focused on demonstrating the significance of data preprocessing, feature selection, and data augmentation in disease prediction that have a significant impact on the model's performance. Six distinct classification algorithms are used in the comparison analysis to highlight the impact with and without Data pre-processing, SMOTE Oversampling.

The work is divided into three sections:

1) With and without data pre-processing, and utilising classification methods to assess accuracy.

2) For feature selection, combining RFE with logistic regression, Random Forest regression, RFECV, Decision tree regression, and Decision tree classifier

3) Using SMOTE oversampling to improve accuracy by decreasing concerns caused by dataset imbalance.

In order to prepare the PIMA and diabetes type data for a deep learning model, data pre-processing is required. The PIMA dataset contains missing values, which means that many attributes have fewer values, such as zeros. The erroneous values reduce the model's performance. The number of null values in each attribute is shown in Table II. Properties including blood pressure, skin thinning, glucose, insulin, and BMI are all zero in Table II, along with other characteristics. Two data pre-processing strategies are used to replace zero values: one removes the zero values and the other replaces the values by finding the mean. Table III shows the results obtained without pre-processing the data. Tables IV and V show the outcomes of two different approaches to data pre-processing. We employ six classifiers in this work for data pre-processing: LR,GNB,KNN,RF,DT, and SVC. When comparing the results, we can see that when the data set is pre-processed, RF achieves an accuracy of 85.7 percent. Several algorithms, including LR, Random Forest Regression, Decision Tree Classifier, Decision Tree Regression, and RFE with cross validation, were used in our work to use the RFE feature selection approach. Following feature selection, the accuracy is determined using five classifiers. When we compare the results of the classifier based on feature selection RFE with Random Forest regression, we find that Random Forest regression has a better outcome DT: 81.16 percent. The third comparison is based on data augmentation with SMOTE Oversampling versus data augmentation without SMOTE Oversampling. The data augmentation method is used to alleviate the issue caused by an unbalanced dataset. Following feature selection, the SMOTE Oversampling technique is used to augment the selected feature dataset in our study. After that, six classifiers are used, and the accuracy of the classifiers is measured. When we examine the results of both methods, with and without SMOTE oversampling, we can see that with SMOTE oversampling, we can attain higher accuracy. Table VIII and Table IX explains the results obtained with and without SMOTE oversampling.

IX. CONCLUSION

The comparison is based on three categories, which are elaborated in the study: (i) with and without data pre-processing, (ii) feature selection using five alternative algorithms, (iii) with and without data augmentation. The importance of data pre-processing, feature selection, and data augmentation can be seen in the three comparisons. When each category is examined separately, data pre-processing comes out on top since it significantly affects how well models or algorithms perform. When we pre-process a dataset, the dataset's quality improves, as does the performance of the models or algorithms. Tables III, IV, and V provide a comprehensive explanation of the contrast. In this study, pre-processing is done in two different ways, and the results of the two methods are compared. We used five different algorithms with RFE and compared the results in Table VII. Likely Feature Selection Importance was also thoroughly explained, and we used five different methods with RFE and compared the results in Table VII. When it comes to data augmentation, the goal is to solve problems that arise from an unbalanced dataset. The work utilised the SMOTE Oversampling technique and conducted a comparison with and without

SMOTE Oversampling, with the findings shown in Tables VIII and IX. From the result obtained it is clear to know the importance of balancing the dataset which improves the performance of the prediction models. Overall, the comparative analysis focus to specify the importance of the data pre-processing which improves the quality of the dataset by fine tuning the inconsistent values with mean value, also the feature selection techniques using the RFE with different classification algorithm helps in finding the suitable attributes that are closely associated with disease and finally data augmentation roles are in disease prediction.

REFERENCES

[1] Khan, R. M. M., Chua, Z. J. Y., Tan, J. C., Yang, Y., Liao, Z., & Zhao, Y. (2019). From pre-diabetes to diabetes: diagnosis, treatments and translational research. Medicina, 55(9), 546.

[2] Nadesh, R. K., & Arivuselvan, K. (2020). Type 2: Diabetes mellitus prediction using deep neural networks classifier. International Journal of Cognitive Computing in Engineering, 1, 55-61.

[3] Rajagopal, A., Jha, S., Alagarsamy, R., Quek, S. G., & Selvachandran, G. (2022). A novel hybrid machine learning framework for the prediction of diabetes with context-customized regularization and prediction procedures. Mathematics and Computers in Simulation, 198, 388-406.

[4] Olisah, C. C., Smith, L., & Smith, M. (2022). Diabetes Mellitus Prediction and Diagnosis from a Data Preprocessing and Machine Learning Perspective. Computer Methods and Programs in Biomedicine, 106773.

[5] Singh, A., Halgamuge, M. N., & Lakshmiganthan, R. (2017). Impact of different data types on classifier performance of random forest, naive bayes, and k-nearest neighbors algorithms.

[6] Maniruzzaman, M., Kumar, N., Abedin, M. M., Islam, M. S., Suri, H. S., El-Baz, A. S., & Suri, J. S. (2017). Comparative approaches for classification of diabetes mellitus data: Machine learning paradigm. Computer methods and programs in biomedicine, 152, 23-34.

[7] García-Ordás, M. T., Benavides, C., Benítez-Andrades, J. A., Alaiz-Moretón, H., & García-Rodríguez, I. (2021). Diabetes detection using deep learning techniques with oversampling and feature augmentation. Computer Methods and Programs in Biomedicine, 202, 105968.

[8] Bukhari, M. M., Alkhamees, B. F., Hussain, S., Gumaei, A., Assiri, A., & Ullah, S. S. (2021). An improved artificial neural network model for effective diabetes prediction. Complexity, 2021.

[9] Roy, K., Ahmad, M., Waqar, K., Priyaah, K., Nebhen, J., Alshamrani, S. S., ... & Ali, I. (2021). An enhanced machine learning framework for type 2 diabetes classification using imbalanced data with missing values. Complexity, 2021.

[10] Khanam, J. J., & Foo, S. Y. (2021). A comparison of machine learning algorithms for diabetes prediction. ICT Express, 7(4), 432-439.

[11] Naz, H., & Ahuja, S. (2020). Deep learning approach for diabetes prediction using PIMA Indian dataset. Journal of Diabetes & Metabolic Disorders, 19(1), 391-403.

[12] Alam, T. M., Iqbal, M. A., Ali, Y., Wahab, A., Ijaz, S., Baig, T. I., ... & Abbas, Z. (2019). A model for early prediction of diabetes. Informatics in Medicine Unlocked, 16, 100204.33333333333333333333.

[13] Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Tang, H. (2018). Predicting diabetes mellitus with machine learning techniques. Frontiers in genetics, 515.

[14] Singh, A., Halgamuge, M. N., & Lakshmiganthan, R. (2017). Impact of different data types on classifier performance of random forest, naive bayes, and k-nearest neighbors algorithms.

[15] Singh, D. A. A. G., Leavline, E. J., & Baig, B. S. (2017). Diabetes prediction using medical data. Journal of Computational Intelligence in Bioinformatics, 10(1), 1-8.

[16] Azrar, A., Ali, Y., Awais, M., & Zaheer, K. (2018). Data mining models comparison for diabetes prediction. Int J Adv Comput Sci Appl, 9(8), 320-323.

[17] Chugh, S., Selvan, K. A., & Nadesh, R. K. (2017, November). Prediction of heart disease using apache spark analysing decision trees and gradient boosting algorithm. In IOP Conference Series: Materials Science and Engineering (Vol. 263, No. 4, p. 042078). IOP Publishing.

[18] Wu, H., Yang, S., Huang, Z., He, J., & Wang, X. (2018). Type 2 diabetes mellitus prediction model based on data mining. Informatics in Medicine Unlocked, 10, 100-107.

[19] Choi, B. G., Rha, S. W., Kim, S. W., Kang, J. H., Park, J. Y., & Noh, Y. K. (2019). Machine learning for the prediction of new-onset diabetes mellitus during 5-year follow-up in non-diabetic patients with cardiovascular risks. Yonsei medical journal, 60(2), 191-199.

[20] Kopitar, L., Kocbek, P., Cilar, L., Sheikh, A., & Stiglic, G. (2020). Early detection of type 2 diabetes mellitus using machine learning-based prediction models. Scientific reports, 10(1), 1-12.

[21] Zhou, H., Myrzashova, R., & Zheng, R. (2020). Diabetes prediction model based on an enhanced deep neural network. EURASIP Journal on Wireless Communications and Networking, 2020(1), 1-13.

[22] I. Gnanadass, "Prediction of Gestational Diabetes by Machine Learning Algorithms," in IEEE Potentials, vol. 39, no. 6, pp. 32-37, Nov.-Dec. 2020, doi: 10.1109/MPOT.2020.3015190.

[23] Reddy, D. J., Mounika, B., Sindhu, S., Reddy, T. P., Reddy, N. S., Sri, G. J., ... & Kora, P. (2020). Predictive machine learning model for early detection and analysis of diabetes. Materials Today: Proceedings.

[24] Singh, N., & Singh, P. (2020). Stacking-based multi-objective evolutionary ensemble framework for prediction of diabetes mellitus. Biocybernetics and Biomedical Engineering, 40(1), 1-22.