International Journal of Advanced Technology and Engineering Exploration, Vol 8(78)
ISSN (Print): 2394-5443 ISSN (Online): 2394-7454
http://dx.doi.org/10.19101/IJATEE.2021.874039

Research Article

Role of attribute selection on tuning the learning performance of Parkinson's data using various intelligent classifiers

K. Alice[1*], Kanimozhi Natesan[2], B. Dhanalakshmi[3] and K. Jaisharma[4]
Associate Professor, Bharath Institute of Higher Education and Research Chennai, India[1]
Assistant Professor, GKM College of Engineering and Technology (Affiliated to Anna University), Chennai, Tamilnadu, India[2]
Professor, Bharath Institute of Higher Education and Research, Chennai, Tamilnadu, India[3]
Assistant Professor, Saveetha School of Engineering, SIMATS, Chennai, Tamilnadu, India[4]

Received: 03-March-2021; Revised: 22-May-2021; Accepted: 25-May-2021

Abstract
Parkinson's Disease (PD) is one of the most common neurodegenerative disorders. It is a chronic disease that reduces dopamine fluid secretion in the brain causes the disorder of both motor and non-motor features. This paper intends to provide a comparative study on the performance measure of various popular machine learning algorithms on the PD dataset obtained from the University of California at Irvine (UCI) machine learning repository. It is observed that biasness prevails in the performance of the classifier towards the majority class due to the imbalanced class distribution of the PD dataset. Hence two most popular preprocessing techniques were employed to balance the dataset one being Synthetic Minority Oversampling TEchnique (SMOTE) and NEAR MISS (NM) an opposite to SMOTE. A SMOTE samples the minority class up to the level of majority class and NM downsamples and brings the majority class down to minority class. All the features in the dataset do not contain useful information about the dataset and also irrelevant data leads to false classification. So, feature reduction is done using information gain ratio and thus obtained reduced dataset is then subjected to classification. For classification five popular classifiers such as Random Forest (RF), Support Vector Machine (SVM), Naive Bayes (NB), K-Nearest Neighbor (KNN) and Decision Trees (DT) were used to compare the performance with the balanced and imbalanced dataset. The evaluation of the classifier's performance is recorded in terms of accuracy, precision, recall, and F-Measure. The results of the conducted experiments show that balancing the majority and minority classes improve precision and recall and there is an increase in accuracy as well as precision. When compared with other classifiers, RF with SMOTE preprocessing was found to be prominent with the information gain greater than 0.18.

Keywords
Parkinson's disease, SMOTE, Near miss, NB, SVM, KNN, DT, RF.

1.Introduction
Dr James Parkinson first defined the Parkinson's Disease (PD) as "shaking palsy" in 1817. PD is one of the most common neurological syndromes of the central nervous system. It is a serious disease targeting aged persons above 60 years of age. PD is identified as a brain disorder [1] that causes the nerve cells to be lost or impaired. These impaired or lost nerve cells will stop the dopamine fluid production in the area of Substantia Nigra. Dopamine is an essential chemical fluid in the brain, which makes nerve cells passes the message to other nerve cells. Due to the lack of dopamine, information has to pass to another nerve cell to be stopped. This leads to the symptoms of Parkinson's disease such as tremor mainly in the hands, limb rigidity, voice impairments, slow movement imbalance and difficulty with walking, writing and talking. Initially, the PD symptoms will begin gradually and it gets worse over time. PD will affect people in various ways. Symptoms are not common to everyone, however voice impairments are the early detected symptoms in PD patients, but cannot be detected by normal listeners [2]. It will get varied based on a person's age, gene, and intensity of the disease. PD was identified as one of the main causes of life threat in the United States of America by the Centers for Disease Control and Prevention. Physicians often use Hoehn and Yahr scale to measure the progression of the disease over the years. This measures the severity of movement symptoms and how much it affects a person's daily life

*Author for correspondence

activities. The scale ranges from zero to five, where zero implies no signs of PD, and scale five indicates the high severity of the disease. PD is identified in more than 10 million people worldwide. Men are affected 1.5 times more than women. Accurately detecting PD at an early stage is very much essential to slow down the progress and also to provide patients with the facility to change the treatment dynamically based on their stage of progression in the disease. Many deep learning techniques are widely used in many health care applications [3, 4] in literature with their own advantages and drawbacks. The main cause behind PD is the lack of ability to repair the dying neurons due to age. Neurochemical fluid dopamine is responsible for sending signals in neuron and for the movement of body parts. As the level of dopamine decreases it slows down the movement of body parts which is not noticed until the condition becomes worse.

World health organization (WHO) record shows that more than 10 million people are affected with PD around the globe [5]. If PD is not detected in the early-stage the disease becomes incurable and result in loss of human lives. The cost of detecting the disease is very high and also the results were not accurate. This leads to the need for the development of automated PD detection systems using machine learning and deep learning techniques. This paper aims to provide a comparative study on the performance measure of various popular machine learning algorithms on the PD dataset obtained from the UCI machine learning repository. This data directed study aimed to evaluate the advancement of Parkinson's disease based on the sternness of the symptoms. Additionally, we aim to balance the PD dataset using Synthetic Minority Oversampling TEchnique (SMOTE) and NEAR MISS (NM) techniques and apply different classifiers to compare and evaluate the performance in classifying balanced and imbalanced PD datasets.

The objectives of this paper can be summarized as

- Feature selection improves the quality of the dataset by eliminating redundant and unrelated features, thus improves the accuracy of the system.
- Imbalances in the dataset impose biases in results and hence balancing the dataset using preprocessing technique yields a correct measurement of classifier performance.
- A comparison study on the performance measure of various popular machine learning algorithms on PD dataset obtained from the UCI machine learning repository.

The remaining of the paper is organized as follows. In section 2, related work is discussed. Section 3 presents methods and materials. The results and discussions are demonstrated in section 4. Finally, the conclusion detailed in section 5.

2.Related work

Ramani et al. [6] surveyed various classification algorithms with the PD dataset. Initially, the dataset was obtained from Max Little from the University of Oxford, in collaboration with the National Centre for Voice and Speech, Denver, Colorado. The input PD dataset consists of vocal measurements of 31 people and among the 31 people 23 were identified with PD. Later on, they have undergone feature relevance analysis with the dataset. The Fisher filtering algorithm found to provide better results. Hence this algorithm is applied to a training dataset with various classification algorithms. Three features were selected based on the results of the fisher filtering algorithm. 13 classification algorithms are taken from a survey with the dataset. Among the various classification algorithms, the Random tree classifier found to be the best classification algorithm with 100% accuracy with zero error rate. The Linear Discriminant Analysis (LDA), Quinlan's C4.5, Decision tree (DT) algorithm (C4.5), Cost-sensitive Decision Tree algorithm (CS-MC4) and K-Nearest Neighbor (KNN) measured 90 % accuracy rate.

Khan [7], applied clustering techniques to get an accurate model for PD prediction. Parkinson's disease dataset has been collected from UCI Repository. Three clustering techniques such as KNN, Random Forest (RF) and AdaBoost were identified to obtain better accuracy results. KNN technique is found to be the best classification model with an accuracy of 90.26%. AdaBoost algorithm is identified as the second-best technique for classification with an accuracy rate of 88.72%, whereas RF obtained the accuracy rate of 87.17%.

Ladha and Pippal et al. [8] performing the clustering with different distance measures to estimate the performance improvement in case of variable distance measures.

Sriram, et al. [9] evaluated the performance of machine learning algorithms to predict PD. PD dataset has been retrieved from the UCI Machine learning repository from the Center for machine learning and Intelligent Systems. The input dataset is composed of vocal measurements of more than 5000 samples analyzing 26 features. The orange tool is used for

report generation in the form of a graph and parallel coordinates. Weka v3.4.10 is used for classification. RF has shown a good accuracy rate of 90.26%, followed by KStar with an accuracy rate of 89.74%

Chahar and Kaur [10] surveyed different machine learning algorithms for the computational analysis and discussion and it is analyzed that these algorithms are found to be useful in different domains and applications [10, 11].

Srinivasan, et al. [12] emphasized various preprocessing techniques, namely Discretization, Resampling, and SMOTE. Artificial Neural Network (ANN) based Multilayer perceptron (MLP) classifier was used for classification purpose in all the experiments. Weka 3.8 tools were used to perform the preprocessing steps in the PD dataset. ANN-based MLP classifier was obtained better accuracy when the dataset is preprocessed with the Discretization and Resampling technique. The combination of SMOTE and Resampling preprocessing techniques on the input set and using MLP classifier leads to higher accuracy.

There are several other methodological implications have been found from other researchers using the machine learning techniques as well as optimization techniques in different domains [13−17, 18].

Gil and Manuel [19] proposed ANN and Support Vector Machine (SVM) with two kernel types to construct classifiers to diagnose Parkinson's disease. The kernel types were used as SVM with linear kernel and SVM with Pearson Universal Kernel (PUK) kernel to check the accuracy rate of the classifier. Parkinson's dataset has been retrieved from the UCI repository. Weka tool was used to experiment with the prediction accuracy with the PD dataset. Furthermore, the confusion matrix with measurable factors such as sensitivity, accuracy, positive and negative predictive values have been created. The proposed method leads to high accuracy rates for sensitivity and negative predictive values with more than 90%.

Khemphila and Boonjing [20] proposed MLP to evaluate the prediction accuracy of the PD dataset with back-propagation learning algorithms. Weka 3.6.6 tool was used to conduct the experiments. Information gain is calculated for 22 attributes. Based on the ranking of information gain, only 16 attributes were selected for classification purposes to reduce the time complexity. Furthermore, ANN was used to classify PD. It has shown the highest accuracy rate in the experimental data set to 91.453% and for the validation data set with 80.769%.

Vásquez-Correa et al. [21] proposed deep learning-based PD based on voice impairments between PD subjects and healthy subjects and validated their experiments using 3 independent datasets with three different languages.

Ali et al. [22] proposed a feature selection based on a PD detection system using multiple voice recordings by taking samples at the same time to improve the PD detection accuracy.

Wang et al. [23] used a deep learning model for automatic discrimination of healthy patients and PD patients based on premotor features.

Lahmiri et al. [24] proposed a PD detection system using SVM based on voice patterns and speech data.

Illner et al. [25] used Sawtooth Waveform Inspired Pitch Estimator (SWIPE) method to analyse the voice pattern disorder caused by PD patients and recorded using a smartphone, provided acceptable results in classifying PD patients from healthy patients but needed better algorithm at low signal to noise ratio level.

Ali et al. [26] developed a PD detection system based on handwritten data. A cascaded learning system Chi2 with Adaboost was developed; in which chi-square is used for feature optimization and Ada booster is used in classification.

Lahmiri and Shmuel [27] focused on PD using pattern recognition, which has eight different varieties of feature selection with nonlinear SVM. Also, they used Bayesian optimization techniques to optimize the kernel functions and classification. The prediction of PD is using 14 voice pattern given reliable result for sensitivity and specificity conditions advantage of this technique. But the computational cost required to predict the PD using nonlinear SVM is very high.

Almeida et al. [28] used phonation audio signals reordered using acoustic cardioid and smartphone devices. Then they implemented 18 feature extraction mechanism and 4 machine learning techniques, namely KNN, MLP, Optimum-Path Forest (OPF) and (SVM. They yield minimum accuracy of 92 % and maximum accuracy of 94% based on the recording device they used in their experimentations. The advantage of their technique is the reliability of the

model using handy smart devices to predict the PD in the early stages. Even though it has some drawback in a larger dimension dataset since it consumes more time for cross-validation testing.

Tracy et al. [29] developed a specially designed vocal biomarker to potentially identify the PD using voice recognition. They extracted the feature using paralinguistic from the voice recordings. Here, their model had the advantage provided to identify the severity levels of PD infected persons. Most of the traditional PD models don't provide the levels of infection. It used to identify the PD in the early stages, even though no symptoms indicated but it is not realistic.

Jain et al. [30] achieved an accuracy of 86.5% by following the strategy of using RF and Gradient Boosting classifiers. They evidenced their significance level by measuring the metrics area under the Receiver Operating Characteristic (ROC) curve (AUC), sensitivity, specificity and accuracy. But the effectiveness of this model is relying on post oversampling. So, the performance of this model is poor in the case of under-sampling collections.

Gunduz [31] utilized the Deep Learning (DL), 9-layer simultaneous feature extraction to predict the PD. To ensure the prediction accuracy they performed Leave-One-Person-Out Cross-Validation techniques. Because of this, their model efficiently predicts the PD even in imbalanced data and it was ensured with F-Measure and Matthews Correlation metrics. Since DL consumes much computational power for parallel feature extraction, this model is not suitable for small computing devices.

The supervised machine learning-based algorithm was proposed by Aich et al. [32]. SVM was clubbed with kernel and they found an accuracy of 95% using the genetic algorithm feature extraction set. Acoustic analysis performed on the recorded voice. They found acoustic clues present in the voice helps in identifying the PD. RF, SVM and NN techniques deployed to identify PD and acoustic analysis was performed. This model supports the linearly separable but does not support machine learning with non-linearly separable.

Ali et al. [33] used a cascading process of using adaptive boosting with Chi2 model analysis. They found the solution for the problem to improve the accuracy to 76% in an imbalanced minority class. Anyway, the accuracy level is needed to be improved a lot in a biased imbalanced model compared to an unbiased model.

Bernardo et al. [34] proposed a PD classification as sick or healthy based on the drawing pattern of the patients. They used OPF, SVM, and naïve Bayes (NB) to process the drawn pattern into 11 feature metrics to analyze the PD. The advantage of their relies on the dedicated software developed for the specific drawing purpose performs well in recording and extracting the features. The disadvantage is that dedicated drawing software currently available for computers, not available for portable devices like smart mobiles, tabulates, etc., Further, the PD early detection also possible by analyzing the patient's Genes,

Peng et al. [35] proposed the algorithm which consists of extracting features, dimensionality reduction and prediction. This model is the combination of ML and DL as well as it takes advantages of both technologies. This becomes the drawback of the system, most of the patient may not have their genes historical records.

The feature selection plays a major role in improving the PD early identification, models like microelectrode recording based model. There are numerous algorithms available like accelerometer-based Gait analysis model, Ground reaction forces model wearable motion sensors-based model, Brain-based classification algorithm, Gait characteristics for classification, etc. [36−42].

3.Methods and materials

3.1 Data set

The PD dataset used in our proposed system had been retrieved from the UCI machine learning repository. The Dataset has 195 records with 24 attributes which were collected from 31 people. 23 out of 31 were affected by Parkinson's disease, and it is represented as 0. The healthy person will be represented as 1 in the status column. PD dataset contains 195 biomedical voice measurements of 31 people. This dataset is imbalanced with 48 samples which are classified as healthy represented with value 0 and 147 samples classified with PD represented with value 1.

In machine learning, we often deal with an imbalanced dataset that consists of samples in one class that will be higher or lower than the other class. To achieve good accuracy, it is expected to maintain an equal number of observations in each class. Most of the machine learning algorithms such as DT, RF, and LR mainly focuses on the majority class. This leads to misclassify the class in various areas such as fault

detection, anomaly detection, and facial recognition. SMOTE and NM-NEAR MISS (Downsampling) algorithms is widely used in machine learning to deal with imbalanced class distribution.

The architecture diagram of the proposed system is as shown in *Figure 1*. Here IG represent the information gain.

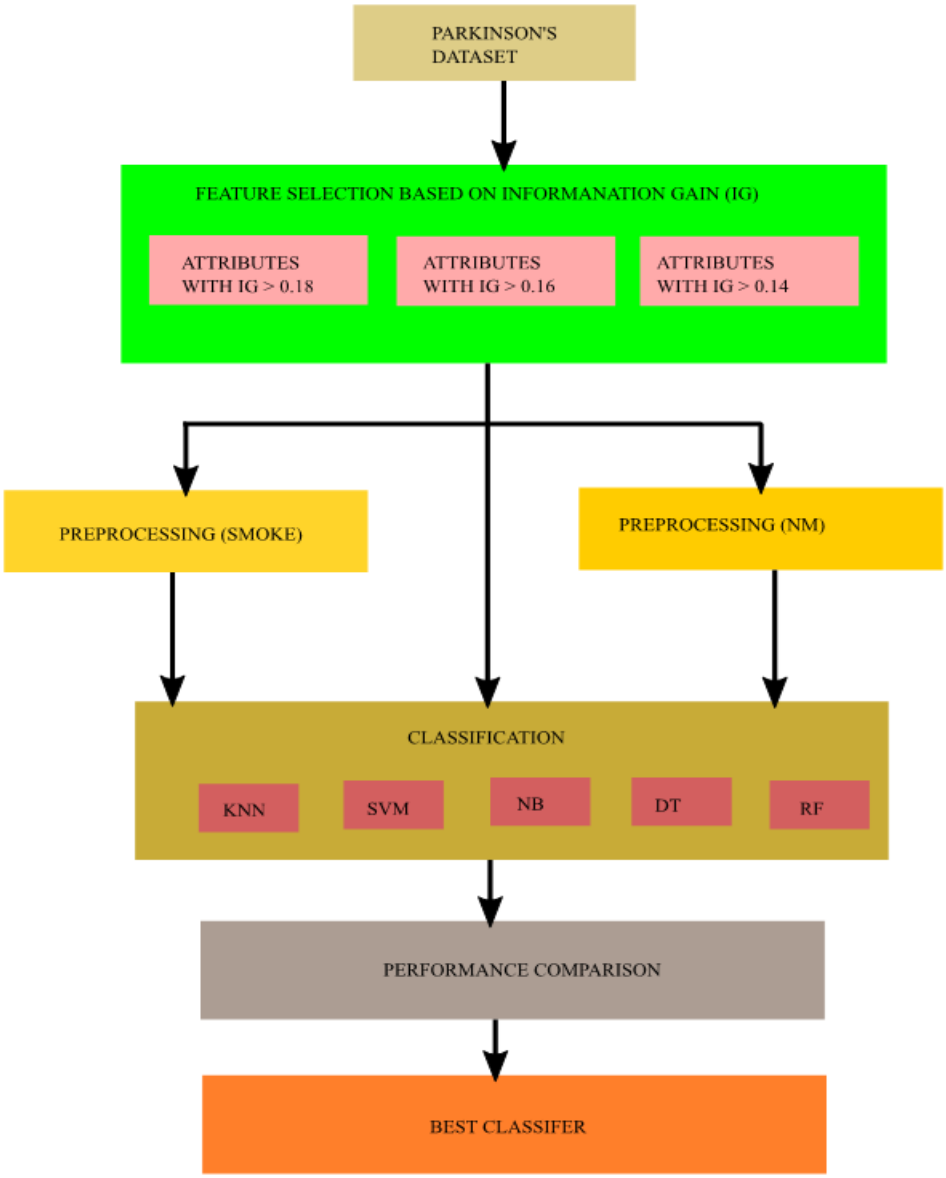

Figure 1 Architecture diagram

3.2 Attribute selection

The dataset is subjected to initial data cleaning operations. PD dataset contains no redundant or null values for attributes. But not all attributes contain useful information about the dataset. So initially all irrelevant information in the dataset is eliminated using attribute selection. Attribute selection is based on the information gain value of each attribute. The information gain values for the input attributes are as shown in *Table 1*. It shows the attributes of the dataset which are biomedical voice measurements of PD patients. For experimental purposes 3 sets of attributes are used. The first set with 11 attributes having an IG greater than 0.18 and the second set with 15 attributes having an IG greater than 0.14 and the third set having 18 attributes with an information gain greater than 0.14.

Table1 Information gain for input attributes

S. No.	Attribute	Description	Gain
1.	Multidimensional Voice Program (MDVP): Flo (Hz)	Minimum vocal fundamental frequency	0.394
2.	Spread1	Nonlinear measure of fundamental frequency variation	0.219
3.	MDVP: Amplitude Perturbation Quotient (APQ)	Measure of variation in fundamental frequency	0.2157
4.	Pitch Period Entropy (PPE)	Nonlinear measure of fundamental frequency variation	0.2103
5.	Noise-to-Harmonic Ratio (NHR)	Measure of ratio of noise to tonal components in the voice	0.1976
6.	spread2	Nonlinear measure of fundamental frequency variation	0.1951
7.	MDVP: Fhi (Hz)	Maximum vocal fundamental frequency	0.1914
8.	MDVP: Relative Average Perturbation (RAP)	Measure of variation in fundamental frequency	0.1881
9.	Jitter: Difference of Differences Perturbation (DDP)	Measure of variation in fundamental frequency	0.1881
10.	MDVP: Shimme	Measure of variation in fundamental frequency	0.1878
11.	Shimmer: Amplitude Perturbation Quotient (APQ5)	Measure of variation in amplitude	0.1828
12.	MDVP: Shimme	Measure of variation in fundamental frequency	0.1753
13.	MDVP: Fo (Hz)	Average vocal fundamental frequency	0.1675
14.	Shimmer: APQ3	Measure of variation in amplitude	0.1609
15.	Shimmer: Difference of Differences between the Amplitudes (DDA)	Measure of variation in amplitude	0.1609
16.	MDVP: Jitter (Abs)	Measure of variation in fundamental frequency	0.1594
17.	MDVP: Point Period Perturbation (PPQ)	Measure of variation in fundamental frequency	0.1565
18.	MDVP: Jitter (%)	Measure of variation in fundamental frequency	0.1484
19.	HNR	Measure of ratio of noise to tonal components in the voice	0.1099
20.	Class	Healthy, Sick	0.1058
21.	Recurrence Period Density Entropy (RPDE)	Nonlinear dynamical complexity measures	0.0844
22.	Correlation Dimension: D2	Nonlinear dynamical complexity measures	0.0783
23.	Detrended Fluctuation Analysis (DFA)	Signal fractal scaling exponent	0.0723

3.3 Preprocessing

SMOTE is an oversampling technique in machine learning. SMOTE will randomly escalate the samples in the minority class. In this oversampling technique, new instances have been generated by synthesizing the value with the existing instances. This will be performed by KNN. It measures the Euclidean distance between the newly generated instances with every other sample in the minority class. In this way, the minority class distribution will get oversampled.

NearMiss is a downsampling technique widely used to handle the imbalanced data distributions in datasets. It aims to balance the class distribution by randomly eliminating the majority class instances. The nearest neighbour algorithm is used to reduce the instances in the majority class also prevent the problem of Information loss in datasets. PD dataset is highly imbalanced with the healthy class being minority class with 48 records and PD class being majority class with 147 records. Imbalance in dataset leads to biased results towards the majority class. So, for balancing the dataset rather than just duplicating the minority class new instances of minority class are synthesized by using the KNN algorithm (k=5) with Euclidean distance metrics. Similarly, majority class instances can also be eliminated up to the range of minority class to achieve balance in the dataset. This can be implemented in python using SMOTE and NM algorithms respectively.

3.4Classification

The two different balanced sets of the dataset along with the actual imbalanced PD dataset are used in

classification. For classification five different classifiers NB with Gaussian NB, SVM with linear Kernel, KNN (k=7), DT and RF were used [6]. The performance of SVM will be relatively good for the smaller dataset as it requires a very high training time for the larger dataset. Also, it performs better for linearly separable binary classifier when a linear kernel is used. KNN classifier is robust against noisy data Ansari and Namdeo [43]. A DT can handle categorical data well and an RF is an ensemble model with a collection of DT working together in the classification. NB is a direct and quick classifier that performs well for unorganized data [44−47]. These widely varying classifiers are used for classification to analyze the performance measures of various classifiers to know the best classifier for the PD dataset.

NB is a supervised learning algorithm, based on the Bayes theorem [44, 45]. It is used for classification problems, mostly in text classification. First, it converts the dataset into frequency values, then it finds the predicted value of has given feature and finally, it uses the Bayes theorem to produce the desired result.

SVM is one of the best-supervised learning algorithms. It is used for regression problems and classification, but it is mostly used for classification problems in machine learning [46]. The SVM algorithm produces the line which can be the best-produced line, according to the mechanism of the algorithm or it can be called the decision boundary line which helps in classifying. In this, the best decision boundary is called a hyper lane. There are linear SVM and non-linear SVM. The boundary line is very useful to categorize the data or given input.

KNN is a machine learning algorithm based on a supervised learning technique. The KNN algorithm uses the stored value or data which can be classified as a new data point based on similarity in the dataset.

DT is one of the most popular machine learning technique. Its input is represented as the branches and nodes in the tree, but the output value is represented as leaf nodes. This technique can be utilized in the classification problem and regression problem. It has a good advantage as it works well with the huge data or it is robust to differentiate the features and it helps us to understand the impact of each variable that is used in the algorithm or calculation. It doesn't work well with the huge training data, thus it led to the poor predictive performance.

RF is a more efficient and widely used classifier. It is highly effective in handling nonlinear classification, especially in the huge dataset. It tries to find out the best node and performs an inbuilt optimization operation [48, 16].

Scikit-learn is a python library that contains a wide range of state-of-the-art machine learning algorithms. The Scikit-learn library was created to help users to apply machine learning algorithms more easily and robust in python.

The experimentation was performed on an HP laptop that has an Intel i7 processor, 16GB of RAM, an NVIDIA 1GB graphics card, and 1TB of a hard disk. The software used to compile the programs using Python 3.8 bundle, Jupyter notebook, Anaconda IDE individual edition installed in Windows 10 operating system. The Parkinson's dataset was downloaded, which is freely available online in the UCI machine learning repository [1]. In the dataset, attributes were selected using information gain of each of the attributes and the same experimental setup is applied to attributes with a gain ratio greater than 0.14, greater than 0.16 and greater than 0.18 with only 18,15 and 11 higher gain ratio attributes are selected out of 23 attributes in the actual PD dataset. The classifiers are compared based on their performance measured in terms of accuracy, precision, recall and F-Score. For performance measure, three different datasets were used: imbalanced PD dataset, the Balanced PD dataset using SMOTE and Balanced PD dataset using NEAR MISS.

The dataset was divided into train and test dataset in the ratio of 70:30. So there are only 136 instances in the training dataset with 101 entries in the majority class, and 35 entries in the minority class. Application of SMOTE on the training dataset improves the dataset with 202 instances in the training set and application NEAR MISS decreases the training set of 70 instances.

For the statistical analysis, IBM SPSS V22 software was used, to find the correlation and coefficient using the mean, standard deviation, standard error, degree factors, significance was analysed by performing ANOVA descriptive analysis [49, 50]]. To find the significance and relationship between groups, Turkey Honest Significance Difference (HSD) Post Hoc test was performed. In our statistical analysis, the sample size was calculated using GPower 3.1 software [49]. The total sample size of 93 was taken divided into 3 groups such as SMOTE 31 samples, NEAR MISS 31 samples and IMBALANCE 31 samples were used.

The confidence interval was taken as 95% and the alpha p-value was taken as 0.05% to evaluate the performance of our model.

4. Results

Figure 2 shows the results of precision, recall and F-Score for class '0' instances and class '1' instances for both imbalanced dataset and balanced (SMOTE AND NM) dataset for attributes with a gain ratio greater than 0.18, greater than 0.16 and greater than 0.14 using NB classifier. There is not much difference in precision, recall and F-Score for both types of dataset taken into consideration. Similar results obtained even with the optimized feature set. It was noted that for the '0' class the recall values are high when compared with precision and the F-Score irrespective of the balanced or imbalanced dataset.

Figure 3 shows the results of precision, recall and F-Scores for class '0' instances and class '1' instances for both imbalanced dataset and balanced (SMOTE AND NM) dataset for attributes with a gain ratio greater than 0.18, greater than 0.16 and greater than 0.14 using SVM with linear kernel classifier. It is clear from the graph that the precision values of the imbalanced dataset for both classes'0' and classes '1' instances are greater than 0.90. This is because of the biasness resulted as a result of imbalance. In a balanced dataset, the results of the precision are reduced for both the classes, but slightly higher for class '0' instances. The recall value of class '0' instances is improved and the recall value of class '1' is decreased on comparing imbalanced dataset with balanced dataset because of the even distribution of instances between classes. It is also noted that the F-Score is decreased in the balanced dataset.

Figure 4 shows the results of precision, recall and F-Score for class '0' instances and class '1' instances for both imbalanced dataset and balanced (SMOTE AND NM) dataset for attributes with a gain ratio greater than 0.18, greater than 0.16 and greater than 0.14 using KKN classifier (k=7). It is clear from the graph that the precision values for class '0' instances of the imbalanced datasets reduced in a balanced dataset using SMOTE by 30 % and NEAR MISS by 60 % and with only a slight reduction (less than 10 %) in class '1' instances. The recall value of class '0' instances is improved and the recall value of class '1' is decreased on comparing imbalanced dataset with balanced dataset because of the even distribution of instances between classes. It is also noted that the F-Score measure decreased in a balanced dataset using the NM algorithm and reported not much difference on the application of SMOTE.

Figure 5 shows the results of precision, recall and F-Score for class '0' instances and class '1' instances for both imbalanced dataset and balanced (SMOTE AND NM) dataset for attributes with a gain ratio greater than 0.18, greater than 0.16 and greater than 0.14 using Decision Tree Classifier. It is clear from the graph that the precision values are decreased for class '0' instances and remains almost the same for class '1' instances.

The recall values are improving for both the class of instances and the F-Score value remains almost the same between different dataset.

Figure 6 shows the results of precision, recall and F-Score for class '0' instances and class '1' instances for both imbalanced dataset and balanced (SMOTE AND NM) dataset for attributes with a gain ratio greater than 0.18, greater than 0.16 and greater than 0.14 using Random Forest Classifier. It is clear from the graph that the precision values are decreased for class '0' instances slightly for SMOTE applied dataset and up to 35 % for the NEAR MISS applied dataset and remains almost the same for class '1' instances. The recall values are improving for class '0' instance to almost 100 % and for class '1' instances it remains unaltered. The F-Score value is improving for SMOTE and decreasing for NEAR MISS when compared with the imbalanced dataset. It was also noted that the results of precision, recall and F-Score is very high in RF when compared with another set of classifiers taken into consideration.

The comparison of SMOTE, NEAR MISS and IMBALANCED, was performed without considering the feature selections as 0.18, 0.16 and 0.14. The result of the same represented in *Table 2*, the variables such as precision, recall, F-Score and accuracy were measured with N=31 samples, mean, standard deviation, standard error were aggregated for the confidence interval 95% and alpha p-value as 0.05%. For the precision variable, the homogeneous subsets have 0.75 for NEAR MISS, 0.80 for SMOTE and 0.88 for IMBALANCED. For the recall variable, the homogeneous subsets have 0.79 for NEAR MISS, 0.87 for SMOTE and IMBALANCED. For the F-Score variable, NEAR MISS have 0.69, SMOTE has 0.81 and IMBALANCE have 0.85. For the accuracy variable, NEAR MISS have 0.73, SMOTE has 0.85 and IMBALANCED have 0.88 respectively. *Table 3* represented using Tukey HSD comparison to find the

significant possibility among the groups. The table shows the results of the mean difference, standard error and significance level for the given hypothesis.

Figure 7 shows the bar chart comparison of precision, recall, F-Score and accuracy performance concerning the group.

The bar chart is drawn by keeping the error bar's value as +/- 1 standard deviation, the error bars plotted on top of each group bars to indicate the probability of type 1 errors in the output result.

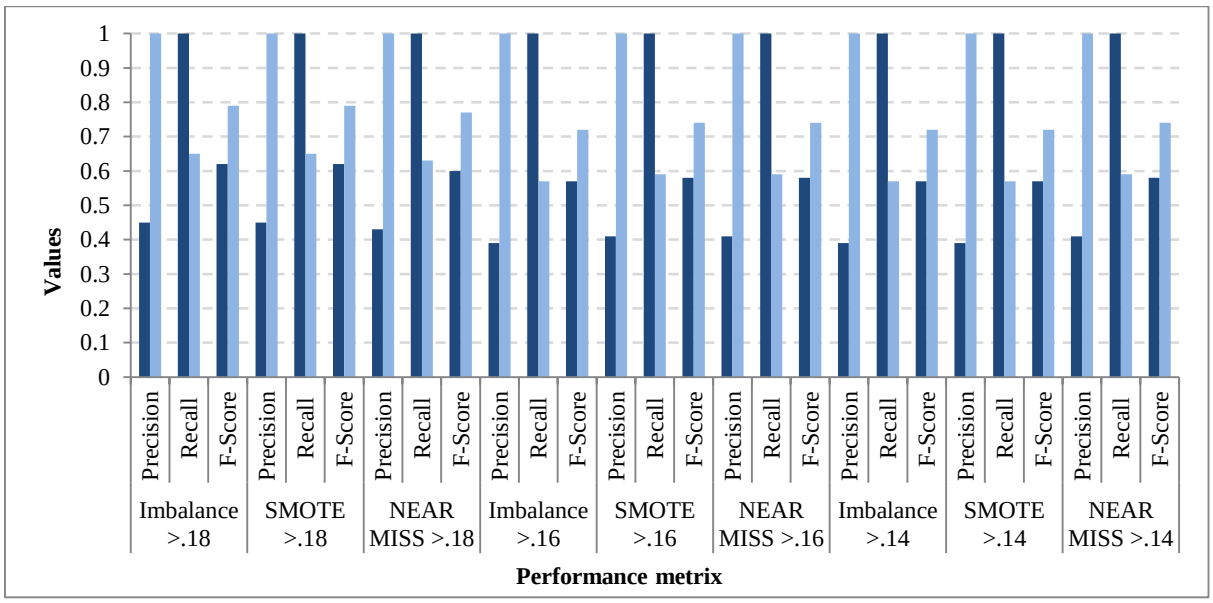

Figure 2 Performance of NB classifier

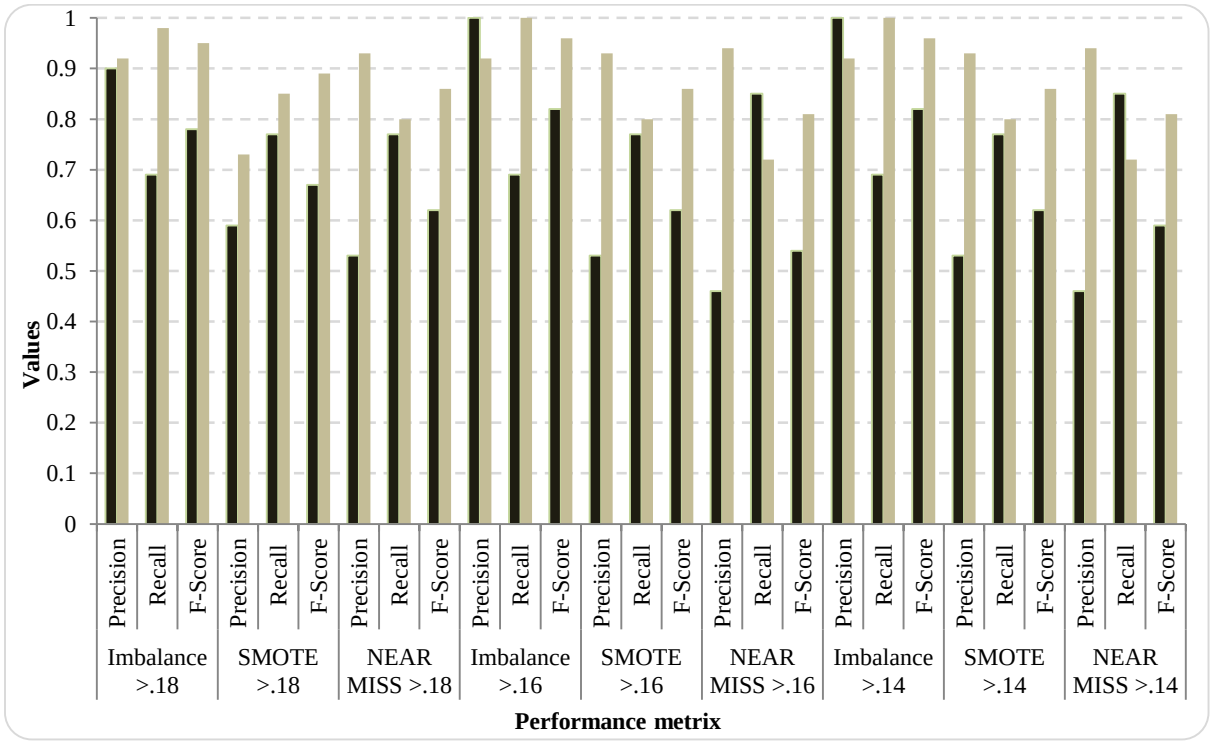

Figure 3 Performance of SVM classifier

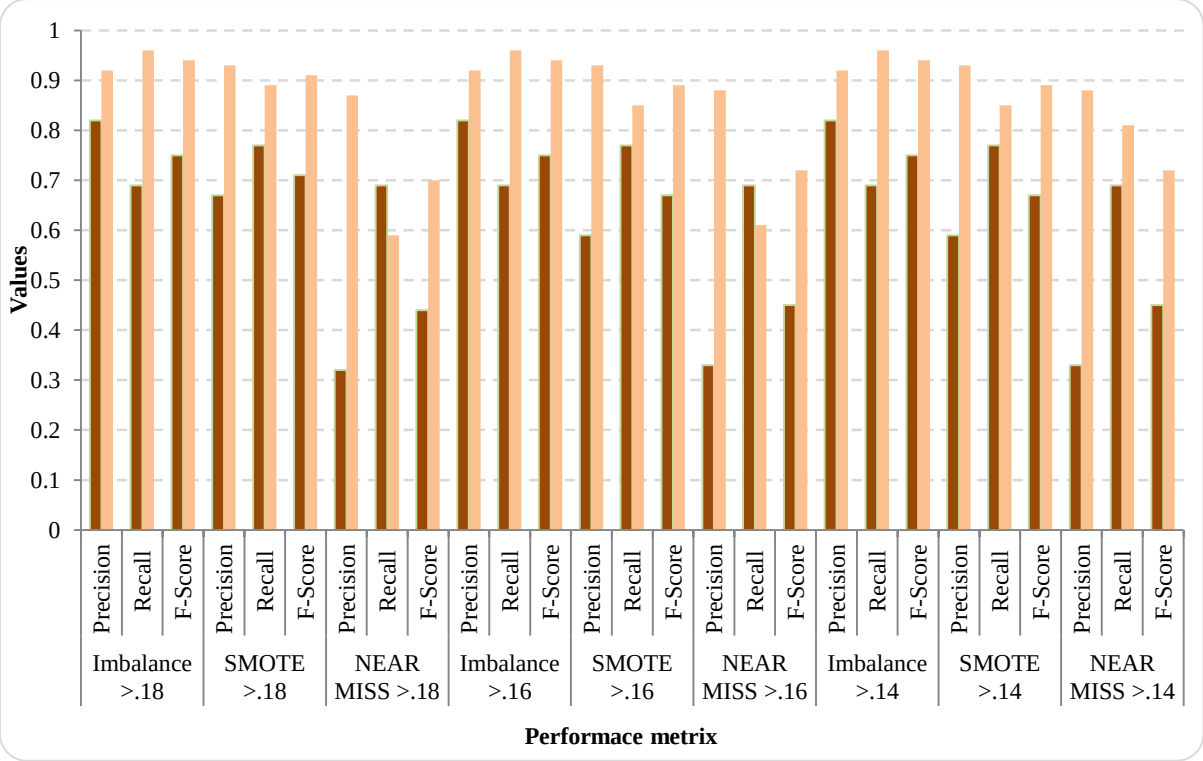

Figure 4 Performance of KNN classifier

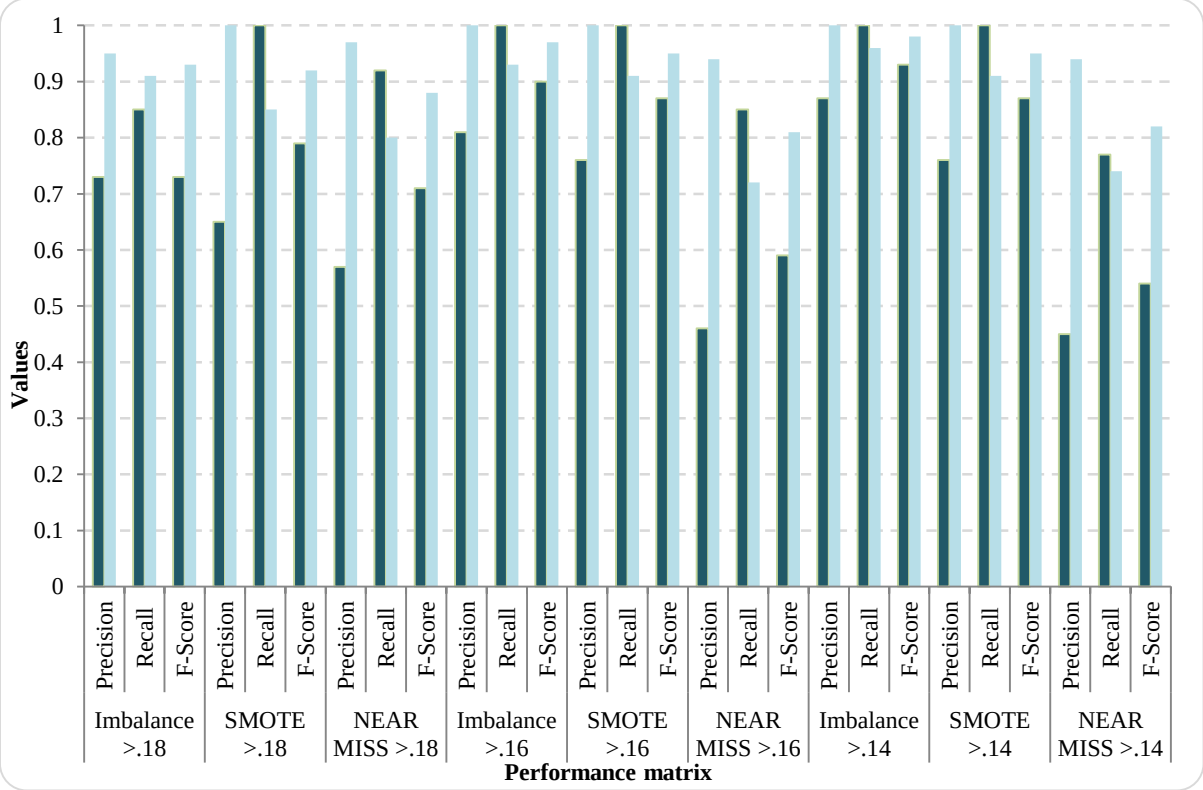

Figure 5 Performance of DT classifier

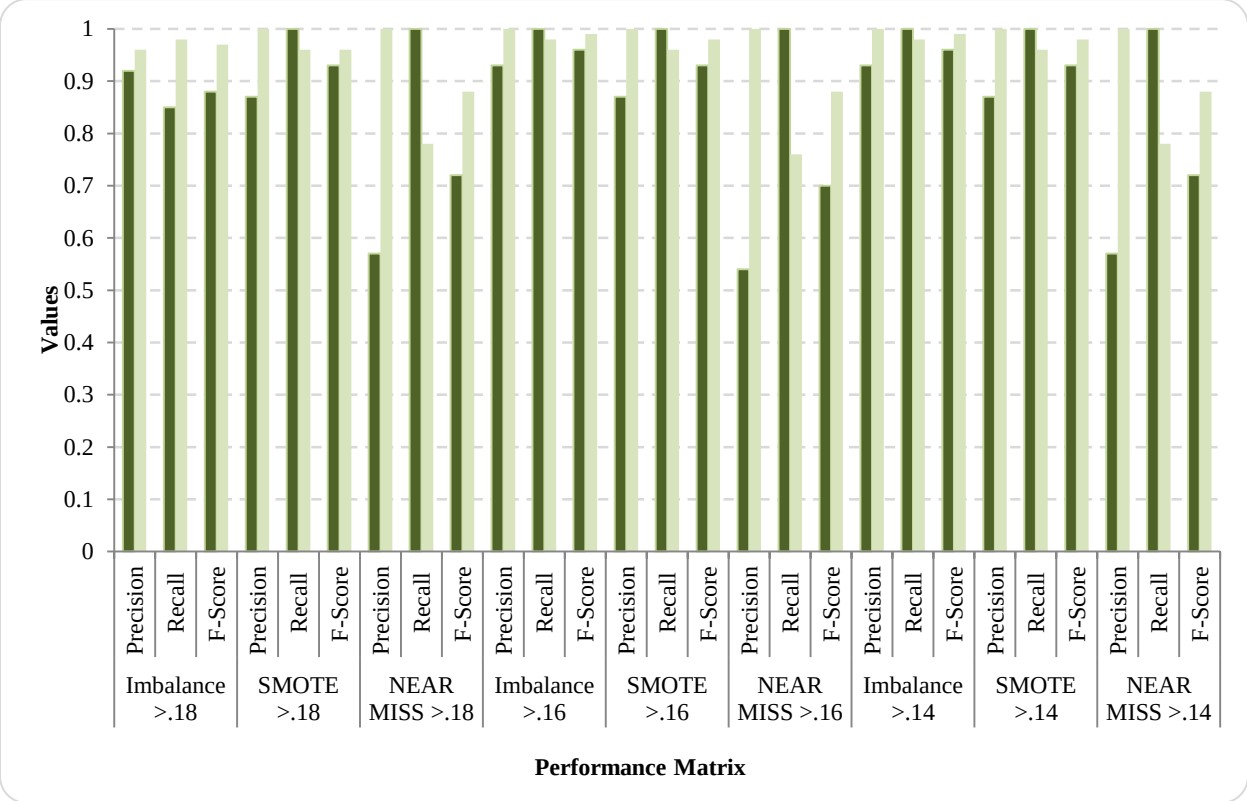

Figure 6 Performance of RF classifier

Table 2 Generalized descriptive analysis of SMOTE, NEAR MISS and IMBALANCED with the variable's precision, recall, F-Score and accuracy without considering the feature selection

		N	Mean	Std. deviation	Std. error	95% Confidence Interval for Mean		Minimum	Maximum
						Lower bound	Upper bound		
Precision	SMOTE	31	.8035	.20552	.03691	.7282	.8789	.39	1.00
	NEAR MISS	31	.7139	.26292	.04722	.6174	.8103	.32	1.00
	IMBALANCE	31	.8777	.17080	.03068	.8151	.9404	.39	1.00
	Total	93	.7984	.22442	.02327	.7522	.8446	.32	1.00
Recall	SMOTE	31	.8729	.12533	.02251	.8269	.9189	.57	1.00
	NEAR MISS	31	.7903	.13167	.02365	.7420	.8386	.59	1.00
	IMBALANCE	31	.8777	.14966	.02688	.8228	.9326	.57	1.00
	Total	93	.8470	.14038	.01456	.8181	.8759	.57	1.00
F-Score	SMOTE	31	.8168	.13479	.02421	.7673	.8662	.57	.98
	NEAR MISS	31	.7010	.13678	.02457	.6508	.7511	.44	.88
	IMBALANCE	31	.8558	.12917	.02320	.8084	.9032	.57	.99
	Total	93	.7912	.14777	.01532	.7607	.8216	.44	.99
Accuracy	SMOTE	31	.8539	.09932	.01784	.8174	.8903	.66	.97
	NEAR MISS	31	.7390	.07591	.01363	.7112	.7669	.61	.83
	IMBALANCE	31	.8871	.10634	.01910	.8481	.9261	.66	.98
	Total	93	.8267	.11336	.01175	.8033	.8500	.61	.98

Table 3 Generalized Turkey HSD comparison of groups SMOTE, NEAR MISS and IMBALANCE with the variable's precision, recall, F-Score and accuracy without considering the feature selection

Dependent variable			Mean difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower bound	Upper bound
Precision	SMOTE	NEAR MISS	.08968	.05497	.238	-.0413	.2207
		IMBALANCE	-.07419	.05497	.372	-.2052	.0568
	NEAR MISS	SMOTE	-.08968	.05497	.238	-.2207	.0413
		IMBALANCE	-.16387*	.05497	.010	-.2949	-.0329
	IMBALANCE	SMOTE	.07419	.05497	.372	-.0568	.2052
		NEAR MISS	.16387*	.05497	.010	.0329	.2949
Recall	SMOTE	NEAR MISS	.08258*	.03453	.049	.0003	.1649
		IMBALANCE	-.00484	.03453	.989	-.0871	.0774
	NEAR MISS	SMOTE	-.08258*	.03453	.049	-.1649	-.0003
		IMBALANCE	-.08742*	.03453	.035	-.1697	-.0051
	IMBALANCE	SMOTE	.00484	.03453	.989	-.0774	.0871
		NEAR MISS	.08742*	.03453	.035	.0051	.1697
F-Score	SMOTE	NEAR MISS	.11581*	.03394	.003	.0349	.1967
		IMBALANCE	-.03903	.03394	.486	-.1199	.0418
	NEAR MISS	SMOTE	-.11581*	.03394	.003	-.1967	-.0349
		IMBALANCE	-.15484*	.03394	.000	-.2357	-.0740
	IMBALANCE	SMOTE	.03903	.03394	.486	-.0418	.1199
		NEAR MISS	.15484*	.03394	.000	.0740	.2357
Accuracy	SMOTE	NEAR MISS	.11484*	.02407	.000	.0575	.1722
		IMBALANCE	-.03323	.02407	.355	-.0906	.0241
	NEAR MISS	SMOTE	-.11484*	.02407	.000	-.1722	-.0575
		IMBALANCE	-.14806*	.02407	.000	-.2054	-.0907
	IMBALANCE	SMOTE	.03323	.02407	.355	-.0241	.0906
		NEAR MISS	.14806*	.02407	.000	.0907	.2054

*. The mean difference is significant at the 0.05 level.

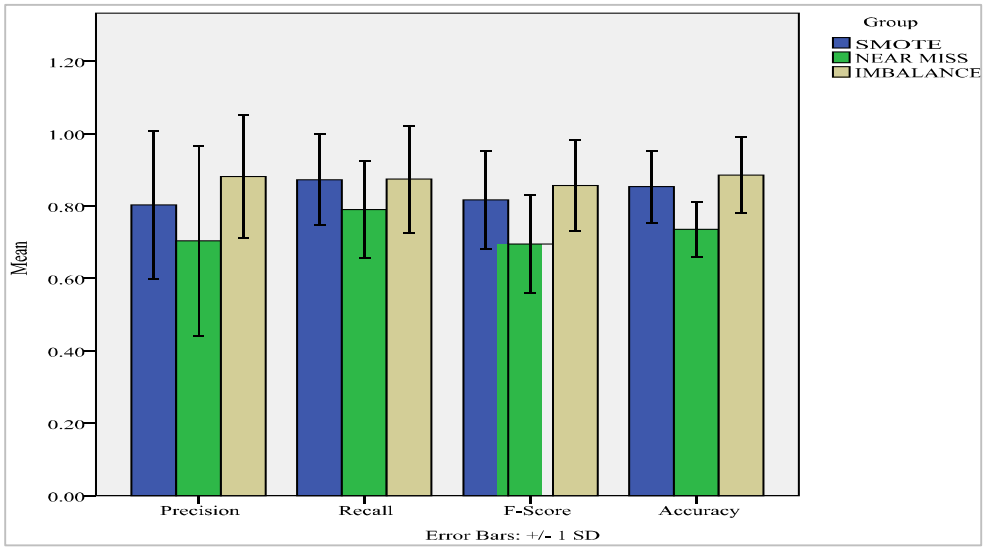

Figure 7 Generalized comparisons of groups SMOTE, NEAR MISS and IMBALANCE for the variable's precision, recall, F-Score and accuracy with the error bars +/- 1 SD

5. Discussion

The data-driven study shows the results of the accuracy of various classifiers for both imbalanced dataset and balanced (SMOTE AND NM) dataset for attributes with a gain ratio greater than 0.18, greater than 0.16 and greater than 0.14. It is clear from the

graph shown in *Figure 8* that the accuracy of the imbalanced dataset is very high when compared to the balanced dataset and is due to the biasness problem caused by majority classes. Even though its accuracy is high it cannot be considered as precise accuracy. The problem with the imbalanced dataset prediction model lies in the classification of the minority class. The model has to learn the features of minority class from very few available samples and that is why the recall of minority class is less in the imbalanced dataset when compared to balanced dataset irrespective of the type of classifier used. The accuracy cause by applying NM is less because of the lesser number of samples of the training dataset with only 70 instances of 35 in each class. It is also noted that the accuracy of RF is high when compared to all other classifiers. It is also noted that the accuracy of all the classifiers is high for the set of attributes whose information gain is greater than 0.16. The limitation of the proposed system is that we have used only five classifiers to study the performance of the classifier in the PD dataset and another classifier like Logistic Regression, Extended Gradient Boost (XGboost), OneR, Multilayer perceptron may also be used which will be considered in a future enhancement.

The study evidenced the use of partially imbalanced biased data, the prediction accuracy can be improved. But it is limited to maintain a sustainable precision score, this will reduce the reliability of the system. Also, this study is conducted using a PD dataset available in the UCI repository, so the result may vary in real-time scenarios. In the near future, this precision score can be improved by concentrating on a different attribute selection and algorithm.

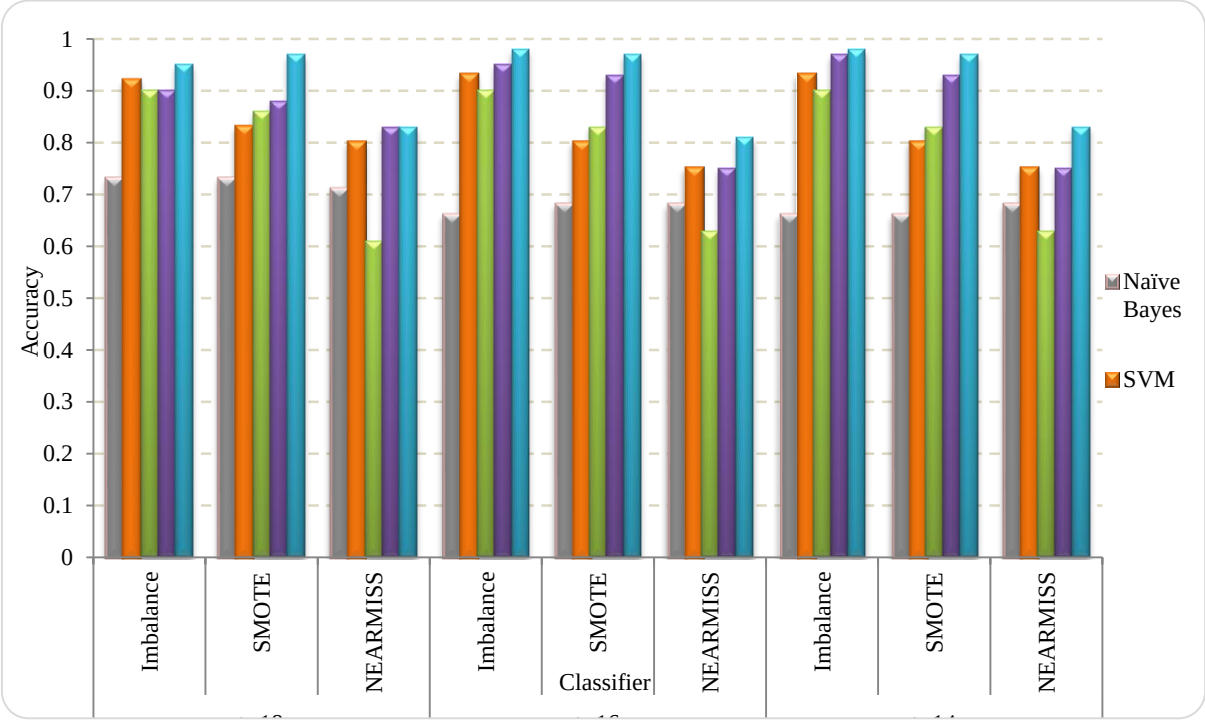

Figure 8 Accuracy for various classifiers

6. Conclusion

This paper shows that the imbalance in the dataset improves the accuracy, but decrease the precision of the system with the help of the PD. Not all features contain useful information about the dataset and irrelevant data may mislead the classifier. A reduction in the dataset is done using information gain and the reduced dataset is used in future processing. The system uses SMOTE and NEAR MISS techniques for oversampling and under-sampling of the dataset. This technique is applied only to the training dataset and the model is built using five different classifiers and test for performance measure using the test dataset. The results reveal that balancing the majority and minority classes improves precision and recall. The recall of minority classes is improved by applying SMOTE and NEAR MISS. The accuracy of the imbalanced dataset is very high when compared to the balanced dataset and is due to the biasness problem caused by majority

classes. When compared with other classifiers taken into consideration the accuracy of RF was found to be prominent for the PD dataset as compared to the other classifiers. It is noted that the accuracy of all the classifiers is high for the set of attributes whose information gain is greater than 0.16. It is also noted that the performance of classifiers for another set of attributes produces similar results. In future, the other feature selection algorithm can be used with the variety of machine learning algorithms.

Acknowledgment
None.

Conflicts of interest
The authors have no conflicts of interest to declare.

References
[1] https://archive.ics.uci.edu/ml/datasets/Parkinsons. Accessed 20 December 2020.
[2] Ali L, Zhu C, Zhang Z, Liu Y. Automated detection of Parkinson's disease based on multiple types of sustained phonations using linear discriminant analysis and genetically optimized neural network. IEEE Journal of Translational Engineering in Health and Medicine. 2019; 7:1-10.
[3] Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, et al. Improved protein structure prediction using potentials from deep learning. Nature. 2020; 577:706-10.
[4] Zhao R, Yan R, Chen Z, Mao K, Wang P, Gao RX. Deep learning and its applications to machine health monitoring. Mechanical Systems and Signal Processing. 2019; 115:213-37.
[5] Johri A, Tripathi A. Parkinson disease detection using deep neural networks. In international conference on contemporary computing 2019 (pp. 1-4). IEEE.
[6] Ramani RG, Sivagami G. Parkinson disease classification using data mining algorithms. International Journal of Computer Applications. 2011; 32(9):17-22.
[7] Khan SU. Classification of Parkinson's disease using data mining techniques. Parkinsons Dis Alzheimer Dis. 2015; 2(1):1-4.
[8] Ladha GG, Pippal RK. An efficient distance estimation and centroid selection based on k-means clustering for small and large dataset. International Journal of Advanced Technology and Engineering Exploration. 2020; 7(73):234-40.
[9] Sriram TV, Rao MV, Narayana GS, Kaladhar DS, Vital TP. Intelligent Parkinson disease prediction using machine learning algorithms. International Journal of Innovative Technology and Exploring Engineering. 2013; 3(3):212-5.
[10] Chahar R, Kaur D. A systematic review of the machine learning algorithms for the computational analysis in different domains. International Journal of Advanced Technology and Engineering Exploration. 2020; 7(71):147-64.
[11] Chahar R. Computational decision support system in healthcare: a review and analysis. International Journal of Advanced Technology and Engineering Exploration. 2021; 8(75):199-220.
[12] Srinivasan SM, Martin M, Tripathi A. ANN based data mining analysis of the Parkinson's disease. International Journal of Computer Applications. 2017; 168(1):1-7.
[13] Mamat RC, Ramli A, Kasa A, Razali SF, Omar MB. Artificial neural networks in slope of road embankment stability applications: a review and future perspectives. International Journal of Advanced Technology and Engineering Exploration. 2021; 8(75):304-19.
[14] Khawaja AW, Kamari NA, Musirin I, Zulkifley MA, Sujod MZ. Design of optimal multi-objective-based facts component with proportional-integral-derivative controller using swarm optimization approach. International Journal of Advanced Technology and Engineering Exploration. 2021; 8(75):391-404.
[15] Salim NA, Jasni J, Mohamad H, Yasin ZM. Transformer health index prediction using feedforward neural network according to scoring and ranking method. International Journal of Advanced Technology and Engineering Exploration. 2021; 8(75):292-303.
[16] Rosdan RM, Awang WS, Bakar WA. Comparison of affinity degree classification with four different classifiers in several data sets. International Journal of Advanced Technology and Engineering Exploration. 2021; 8(75):247-57.
[17] Saikia D, Boruah PK, Sarma U. A computational model for optimum process parameters based on factory data and overall liquor rating of black tea. International Journal of Advanced Technology and Engineering Exploration. 2020; 7(73):220-33.
[18] Braga D, Madureira AM, Coelho L, Ajith R. Automatic detection of Parkinson's disease based on acoustic analysis of speech. Engineering Applications of Artificial Intelligence. 2019; 77:148-58.
[19] Gil D, Manuel DJ. Diagnosing parkinson by using artificial neural networks and support vector machines. Global Journal of Computer Science and Technology. 2009; 9(4):63-71.
[20] Khemphila A, Boonjing V. Parkinsons disease classification using neural network and feature selection. International Journal of Mathematical and Computational Sciences. 2012; 6(4):377-80.
[21] Vásquez-Correa JC, Arias-Vergara T, Orozco-Arroyave JR, Eskofier B, Klucken J, Nöth E. Multimodal assessment of Parkinson's disease: a deep learning approach. IEEE Journal of Biomedical and Health Informatics. 2018; 23(4):1618-30.
[22] Ali L, Zhu C, Zhou M, Liu Y. Early diagnosis of Parkinson's disease from multiple voice recordings by simultaneous sample and feature selection. Expert Systems with Applications. 2019; 137:22-8.
[23] Wang W, Lee J, Harrou F, Sun Y. Early detection of Parkinson's disease using deep learning and machine learning. IEEE Access. 2020; 8:147635-46.
[24] Lahmiri S, Dawson DA, Shmuel A. Performance of machine learning methods in diagnosing Parkinson's

disease based on dysphonia measures. Biomedical Engineering Letters. 2018; 8(1):29-39.
[25] Illner V, Sovka P, Rusz J. Validation of freely-available pitch detection algorithms across various noise levels in assessing speech captured by smartphone in Parkinson's disease. Biomedical Signal Processing and Control. 2020.
[26] Ali L, Zhu C, Golilarz NA, Javeed A, Zhou M, Liu Y. Reliable Parkinson's disease detection by analyzing handwritten drawings: construction of an unbiased cascaded learning system based on feature selection and adaptive boosting model. IEEE Access. 2019; 7:116480-9.
[27] Lahmiri S, Shmuel A. Detection of Parkinson's disease based on voice patterns ranking and optimized support vector machine. Biomedical Signal Processing and Control. 2019; 49:427-33.
[28] Almeida JS, Rebouças FPP, Carneiro T, Wei W, Damaševičius R, Maskeliūnas R, et al. Detecting Parkinson's disease with sustained phonation and speech signals using machine learning techniques. Pattern Recognition Letters. 2019; 125:55-62.
[29] Tracy JM, Özkanca Y, Atkins DC, Ghomi RH. Investigating voice as a biomarker: deep phenotyping methods for early detection of Parkinson's disease. Journal of Biomedical Informatics. 2020; 104:1-10.
[30] Jain D, Mishra AK, Das SK. Machine learning based automatic prediction of Parkinson's disease using speech features. In proceedings of international conference on artificial intelligence and applications 2021 (pp. 351-62). Springer, Singapore.
[31] Gunduz H. Deep learning-based Parkinson's disease classification using vocal feature sets. IEEE Access. 2019; 7:115540-51.
[32] Aich S, Kim HC, Hui KL, Al-Absi AA, Sain M. A supervised machine learning approach using different feature selection techniques on voice datasets for prediction of Parkinson's disease. In international conference on advanced communication technology 2019 (pp. 1116-21). IEEE.
[33] Ali L, Wajahat I, Golilarz NA, Keshtkar F, Bukhari SA. LDA–GA–SVM: improved hepatocellular carcinoma prediction through dimensionality reduction and genetically optimized support vector machine. Neural Computing and Applications. 2021; 33(7):2783-92.
[34] Bernardo LS, Quezada A, Munoz R, Maia FM, Pereira CR, Wu W, et al. Handwritten pattern recognition for early Parkinson's disease diagnosis. Pattern Recognition Letters. 2019; 125:78-84.
[35] Peng J, Guan J, Shang X. Predicting Parkinson's disease genes based on node2vec and autoencoder. Frontiers in Genetics. 2019; 10:1-6.
[36] Wan KR, Maszczyk T, See AA, Dauwels J, King NK. A review on microelectrode recording selection of features for machine learning in deep brain stimulation surgery for Parkinson's disease. Clinical Neurophysiology. 2019; 130(1):145-54.
[37] Rastegari E, Azizian S, Ali H. Machine learning and similarity network approaches to support automatic classification of Parkinson's diseases using accelerometer-based gait analysis. In proceedings of the Hawaii international conference on system sciences 2019(pp.4231-42).
[38] Veeraragavan S, Gopalai AA, Gouwanda D, Ahmad SA. Parkinson's disease diagnosis and severity assessment using ground reaction forces and neural networks. Frontiers in Physiology. 2020; 11:1-11.
[39] Moon S, Song HJ, Sharma VD, Lyons KE, Pahwa R, Akinwuntan AE, et al. Classification of Parkinson's disease and essential tremor based on balance and gait characteristics from wearable motion sensors via machine learning techniques: a data-driven approach. Journal of NeuroEngineering and Rehabilitation. 2020; 17:1-8.
[40] Olivares R, Munoz R, Soto R, Crawford B, Cárdenas D, Ponce A, et al. An optimized brain-based algorithm for classifying Parkinson's disease. Applied Sciences. 2020; 10(5):1-16.
[41] Ricciardi C, Amboni M, De Santis C, Ricciardelli G, Improta G, Iuppariello L, et al. Classifying different stages of Parkinson's disease through random forests. In mediterranean conference on medical and biological engineering and computing 2019 (pp. 1155-62). Springer, Cham.
[42] Rehman RZ, Del Din S, Guan Y, Yarnall AJ, Shi JQ, Rochester L. Selecting clinically relevant gait characteristics for classification of early Parkinson's disease: a comprehensive machine learning approach. Scientific Reports. 2019; 9:1-12.
[43] Ansari HF, Namdeo V. An efficient SKNN based approach for heart disease classification. International Journal of Advanced Technology and Engineering Exploration. 2019; 6(53):101-6.
[44] Granik M, Mesyura V. Fake news detection using naive bayes classifier. In first Ukraine conference on electrical and computer engineering 2017 (pp. 900-3). IEEE.
[45] Bužić D, Dobša J. Lyrics classification using naive bayes. In international convention on information and communication technology, electronics and microelectronics 2018 (pp. 1011-5). IEEE.
[46] Copur M, Ozyildirim BM, Ibrikci T. Image classification of aerial images using CNN-SVM. In innovations in intelligent systems and applications conference 2018 (pp. 1-6). IEEE.
[47] Vera JE, Martinez SM, Pérez AT, Avendano J. Classification of gerbera type flowers based in decision tree rules. In symposium on image, signal processing and artificial vision 2019 (pp. 1-4). IEEE.
[48] Paul A, Mukherjee DP, Das P, Gangopadhyay A, Chintha AR, Kundu S. Improved random forest for classification. IEEE Transactions on Image Processing. 2018; 27(8):4012-24.
[49] O'Neill E, Yssel JD, McNamara C, Harkin A. Pharmacological targeting of β2-adrenoceptors is neuroprotective in the LPS inflammatory rat model of Parkinson's disease. British Journal of Pharmacology. 2020; 177(2):282-97.
[50] Carrilho PE, Rodrigues MA, Oliveira BC, Silva EB, Silva TA, Schran LD, et al. Profile of caregivers of

Parkinson's disease patients and burden measured by zarit scale analysis of potential burden-generating factors and their correlation with disease severity. Dementia & Neuropsychologia. 2018; 12(3):299-305.

K. Alice received her PhD Degree in Information and Communication Engineering from Anna University, India in 2020, and the M.E Degree in Computer Science and Engineering from Sathyabama University, India in 2005 and the B.E Degree in Computer Science and Engineering from Madurai Kamaraj University, India in 1999. She is currently working as an Associate Professor at Bharath Institute of Higher Education and Research Chennai, India. Her research interests are in Image Processing, Deep Learning and Information Security.
Email: k_alice_suresh@yahoo.com

Kanimozhi Natesan received her B.TECH in Information Technology from the Annai Mathammal Sheela Engineering College, Namakkal in 2008 and M.E. in Computer Science and Engineering at the Annai Mathammal Sheela Engineering College, Namakkal in 2010. She is pursuing her PhD in Information and Communication at Anna University, Chennai. She is currently working as an Assistant Professor in the Department of Computer Science and Engineering at GKM College of Engineering and Technology (Affiliated to Anna University), Chennai, Tamilnadu, India. She has 8 years of teaching experience and three years in research. Her specializations include Network Security, Cloud Computing, Internet of Things, Deep Learning and Image Processing. Her current research interests are Deep Learning and Machine Learning.
Email: kanimozhiraja3108@gmail.com

B. Dhanalakshmi received her Bachelor's degree in Information Technology from Sathyabama Institute of Science and Technology, Master's degree in Information Technology from Sathyabama University. She also has completed her Degree of Doctor of Philosophy under the Faculty of Computer Science and Engineering from Sathyabama Institute of Science and Technology. She is currently working as a Professor in the Department of Information Technology in Bharath Insitute of Higher Education and Research, Chennai, Tamilnadu, India and her area of interest include Data Mining, Artificial Intelligence, Object-Oriented Analysis and Design, Computer Architecture and Software Engineering.
Email: dhana.baskaran@gmail.com

K. Jaisharma received his Bachelor Degree in Computer Science and Engineering from Vel Tech Engineering College, Chennai and M.Tech in Information Technology from Bharath University, C. He is pursuing his PhD in Deep Learning at Saveetha University, Chennai. He is currently working as an Assistant Professor in the Department of Computer Science and Engineering at Saveetha School of Engineering, SIMATS, Chennai, Tamilnadu, India. He has 8 years of teaching experience and his area of interest include Machine Learning, Deep Learning, Internet of Things, Data Analytics and Mobile Application Development
Email: k.jaisharma@gmail.com