Machine Learning Model to Analyze Telemonitoring Dyphosia Factors of Parkinson's Disease

Mohimenol Islam Fahim[1], Syful Islam[2],
Sumaiya Tun Noor[3], Md. Javed Hossain*[4]
Department of Computer Science and Telecommunication Engineering
Noakhali Science and Technology University
Noakhali-3814, Bangladesh

Md. Shahriar Setu[5]
Department of Management Information Systems
Noakhali Science and Technology University
Noakhali-3814, Bangladesh

Abstract—**For many years, lots of people have been suffering from Parkinson's disease all over the world, and some datasets are generated by recording important PD features for reliable decision-making diagnostics. But a dataset can contain correlated data points and outliers that can affect the dataset's output. In this work, a framework is proposed where the performance of an original dataset is compared to the performance of its reduced version after removing correlated features and outliers. The dataset is collected from UCI Machine Learning Repository, and many machine learning (ML) classifiers are used to evaluate its performance in various categories. The same process is repeated on the reduced dataset, and some improvement in prediction accuracy is noticed. Among ANOVA F-test, RFE, MIFS, and CSFS methods, the Logistic Regression classifier along with RFE-based feature selection technique outperforms all other classifiers. We observed that our improved system demonstrates 82.94% accuracy, 82.74% ROC, 82.9% F-measure, along with 17.46% false positive rate and 17.05% false negative rate, which are better compared to the primary dataset prediction accuracy metric values. Therefore, we hope that this model can be beneficial for physicians to diagnose PD more explicitly.**

Keywords*—*Parkinson's disease; correlation; outliers; machine learning; RFE-based analysis

I. Introduction

Parkinson's disease (PD) is a chronic, neurodegenerative disease of the nervous system which affects our body movement including speech [1]. James Parkinson was invented this disease in 1857 and explained its condition as Shaking Palsy [2]. The main reason of PD is actually unknown. It affects 1% of people who are older than 65 years, and no medical treatments can cure this disease completely [3]. Almost 90% patients face trouble speaking normally as well as fail to express facial emotion; it results in slow speaking speed, slur words, mumbling, etc. [4]. The average age of patients lies between 55 to 65 years old [5]. Different environmental factors like rural living, consumption of water, pesticide manage and exposure, environmental toxin create individual's risks of happening PD. Out of many neurodegenerative disease such as Alzheimer's disease, headache disorders, stroke, epilepsy, multiple sclerosis, dementia, PD is considered as the second most common nerudegenerative disorder [2]. Different brain cells contain substantia nigra cells which produce dopamine. Dopamine is a chemical element which transmits signals within brain and controls the movement of body. When 60-80% dopamine creating cells are lost, there are not produced sufficient dopamine and people face about movement disorder that causes PD [6].

To ensure proper treatment about PD, it is required to identify these patients as early as possible. Many works have been happened where PD patients are identified based on different aspects and parameters. The symptoms of PD is divided into motor and non-motor group. The motor group is also called as cardinal symptoms which include tremor, rigidity, postural instability, and slowness of movement. Instead, non-motor group shows the loss of speech, facial expression, and handwriting. These types of symptoms are called dopamine non-responsive symptoms. Speech properties are one of the most effective non motor element because 90% patients are faced PD based on vocal impairment [7]. In addition, non motor symptoms like speech are not decisive where these attributes are employed with cerebrospinal fluid measurement (CSF) and dopamine transporter imaging for predicting PD [8]. Due to redundant points and degradation of speech quality, it is more difficult for physicians to detect PD cases by assessing their vocal records in a manual way. Thus, an automatic model is useful which extracts speech patterns of subjects and detects PD more efficiently.

However, machine learning is a study of computer algorithms where it analyzes existing instances and predict expected outcomes [9], [10]. It is defined as a process of discovering useful, interesting, and complex patterns from a large amount and high dimensional data [11], [12]. Likewise, this technique is useful to predict PD through a set of practical datasets. In this work, we propose a machine learning-based framework to make PD detection convenient for clinicians. This model contains various state-of-art techniques like feature selection, outlier detection, and classification. Then, several evaluation metrics like accuracy, area under curve (AUC), f-measure, g-mean, sensitivity, specificity, fall-out, and miss rate are used to assess the performance of individual classifiers [13]. The performance of classifiers are useful to detect the most significant feature subset where different classifier performs well than other subsets. The main contributions of this proposed PD diagnosis model are mentioned below:

- Various feature subsets are generated and identified the best one by assessing the performance of individual classifiers.
- Detect anomalous/noisy elements to obtain more suitable feature subsets.

- To justify the performance of classifiers, numerous evaluation metrics are considered in this work.

This paper is organized as follows: Section 2 includes details of similar studies and their implications. Section 3 presents the methodology of a machine learning model for detecting PD at early stage. Also, it outlines the description of PD dataset, feature selection, classification and its evaluation metrics. Section 4 shows the experimental results of various classifiers for individual feature subdatasets, compare them to identify best feature subset. Finally, Section 5 concludes by summarizing this work and mentioning future research strategies.

II. Related Work

Numerous works were happened to predict PD at early stage. Das [14] used different classifiers like Artificial Neural Network (ANN), DMneural, Regression, and Decision Tree (DT) to efficiently detect PD and compare their results. Tsana et al. [15] employed novel speech signal processing feature selection and statistical classifiers to investigate PD. Challa et al. [8] developed an automatic PD diagnosis model with feature extraction and various classifiers such as Multilayer Perceptron (MLP), Bayes Net (BN), Random Forest (RF), and boosted LR for early prediction of PD. Shamli et al. [16] proposed a multi-class classification model including C4.5, Support Vector Machine (SVM), and ANN to enhance prediction tendencies as well as reduce the cost for PD. Tong et al. [17] proposed a machine learning framework that achieves a 75% classification accuracy along with 69% balanced accuracy for neurodegenerative disease diagnosis. Since PD is a neurodegenerative disease as well, their system can improve the prediction rate for clinical use. Li et al. [18] proposed a PD-oriented classification algorithm for improved classification performance. It involves a Classification and Regression Tree (CART) approach for picking the optimal training samples iteratively and an ensemble-learning algorithm combining RF, SVM, and ELM. Mathur et al. [19] implemented various classifiers like SMO, KNN, Rf, AdaBoost.MI, Bagging, MLP, and DT to scrutinized PD. Nilashi et al. [5] proposed a hybrid intelligent system for PD prediction where Incremental SVM is utilized to estimate Total-UPDRS and Motor-UPDRS. Almeida et al. [20] used 18 feature extraction and 4 machine learning methods to investigate sustainable phonation and speech tasks. Besides, phonation analysis was more efficient than speech task. Lahmini and Shmuel [21] investigated PD based voice pattern using various pattern ranking methods and optimized SVM. Mostafa et al. [22] proposed a new multiple feature evaluation approach (MFEA) as well as DT, NB, ANN, RF, and SVM show its best results for MFEA. Pham et al. [7] combined voice and image dataset where pairwise correlation and k-means clustering extracts features from vocal dataset. Then, it proposed an ensemble method to predict PD. Pahuja et al. [2] extracted various significant features and selected feature subsets from PD voice input dataset. Then, different classifiers such as ANN, SVM, and KNN were implemented and ANN with levenberg-marquardt algorithm provides the best results. Senturk et al. [6] proposed a machine learning model where feature importance and recursive feature elimination (RFE) methods were implemented for feature selection. Then, CART, ANN, and SVM were used to identify PD patients. Karabayir et al. [23] analyzed PD acoustic data using light and extreme gradient boosting, RF, SVM, KNN, LASSO, and LR. Then, they used feature importance procedure to identify significant features for classifying PD. Lamba et al. [24] represented a speech signal based hybrid PD disease diagnosis system where numerous feature selection (i.e., mutual information gain, extra tree, genetic algorithm) and classification methods (i.e., NB, KNN, RF) were employed. Also, SMOTE method was used to balance PD dataset. Paramanik et al. [25] used two recent decision forest algorithms such as SysFor, ForestPA including RF for developing PD detection models with the optimization of DT.

III. Materials and Methods

In this work, we propose a machine learning framework to improve the efficiency of a PD dataset where the data validity is judged by applying many classifiers. For each classifier, multiple performance parameters are measured where we observed that these results could be improved by removing insignificant features and outliers. In the feature selection process, we employ a total of four methods and notice its outcomes.

A. Parkinson's Disease Data

We collected the dataset from the University of California Irvine (UCI) Machine Learning Repository, approved by the Bioethical Committee from the University of Extremadura. The dataset was created by Naranjo et al. [26]. It contains 240 instances for only 80 people whose ages are greater than 50 years old. Among 40 controls, there are found 22 men and 18 women respectively. On the other hand, 27 men and 13 women are defined as PD patients. According to the mean of Unified Parkinson's Disease Rating Scale (UPDRS), all subjects have 5 years or less PD duration. This dataset contains 44 acoustic features which captures a sustainable vowel /a/ for 5s with three runs. These features include five categories such as pitch local features, amplitude local perturbation, special envelope, noise and nonlinear measures. The individual features of these categories are given as follows:

- **Pitch Local Features:** jitter relative, jitter absolute, jitter relative absolute perturbation (RAP), jitter pitch perturbation quotient (PPQ).
- **Amplitude Perturbation Measures:** shimmer local, shimmer dB, 3 point amplitude perturbation quotient (APQ3), 5 point Amplitude Perturbation Quotient (APQ5), 11 point Amplitude Perturbation Quotient (APQ11).
- **Noise:** Harmonic-to-Noise Ratio (HNR) such as HNR05 [0–500 Hz], HNR15 [0–1500 Hz], HNR25 [0–2500 Hz], HNR35 [0–3500 Hz], HNR38 [0–3800 Hz], Glottalto-Noise Excitation Ratio (GNE).
- **Special Envelope:** 13 Mel Frequency Cepstral Coefficients (MFCCs) and 13 Delta Coefficients.
- **Non Linear Measure:** Recurrence Period Density Entropy (RPDE), Detrended Fluctuation Analysis (DFA), and Pitch Period Density Entropy (PPE).

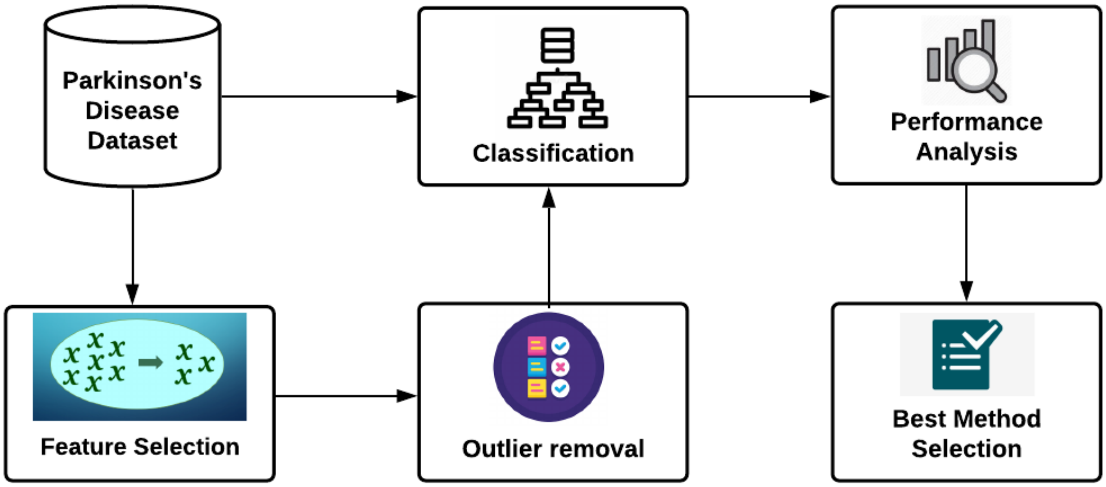

Fig. 1. Pipeline Diagram of the Overall Methodology.

B. Methodology

The he overall implementation are demonstrated in the following Fig. 1:

1) Data Acquisition: After gathering PD voice dataset for UCI data repository, we clean and check missing, wrong, and incomplete information in this dataset. Afterwards, this dataset is prepared for further analysis.

C. Feature Selection Methods

Feature selection methods are useful to reduce the number of input variables and lessen the computational cost of these predictive models. In this work, we apply different feature selection methods into primary dataset and explore several feature subsets. Then, some sub datasets are generated using these subsets.

1) Correlation based Feature Selection (CFS): Correlated values are linearly dependent on each other. Some features don't have any significant impact on the predicted responses, but they have a few drawbacks. A correlation matrix is created to find out the correlation among different features and remove some of them have higher coefficients above a particular limit [27]. It is a square matrix that consists of equal dimensions as features where all the possible correlated pairs are identified and displayed altogether. In order to drop them, a threshold is considered so that all columns exceeding this limit are eliminated. As expected, the number of columns of our dataset is decreased now, and it only contains features having a coefficient less than 0.90.

2) Analysis of Variance (ANOVA) F-test: ANOVA F-test [28] is really helpful to determine if more than one data samples' mean can be driven from the same or different distribution. On the other hand, F-statistic or F-test refers to a class of statistical tests, where the ratio between variances are measured. ANOVA F-test method can be applied to detect the most important features to minimize high data dimensionality. It is a common feature selection strategy for numerical input values and categorical target variables.

3) Chi-Square Feature Selection (CSFS): CSFS is used to evaluate the discrepancy from the expected distribution when the feature incidence is independent from class value [29]. It tests two individual examples to avoid overfitting, reduce computational time, and boost the system's accuracy. However, it can work with data values measured on a nominal scale. The differences between various participant groups can be easily estimated without any assumptions about the distribution.

4) Mutual Information based Feature Selection (MIFS): MIFS represents statistical independence that determines the relationships between random variables [30]. In brief, it detects the quantity of information one random value contains about another one. When it is used as a feature selection scheme, it gives the model a chance to evaluate the relevance of feature subsets depending on the output vector. By quantifying the gain, the system can make effective feature selection decisions.

5) Recursive Feature Elimination (RFE): RFE [31], [32] is effective at picking more relevant parameters in large training datasets. While using RFE, programmers should pay full attention to the number of features selection and the right algorithm implementation. It operates by looking for a subset of features for all columns of the training dataset and getting rid of some irrelevant features. At first, the classifier gets trained, and parameters whose absolute values are the smallest get eliminated until only the required ones remain.

D. Outlier Detection

Outliers refer to those data points, whose have a significant difference from common observations, for the variability of measurement, sampling issues, and experimental errors [33]. These values deviate outcomes from expected values in further analysis. So, we simply address them as deviant examples, unusual data, and special samples respectively. In many cases, they do not provide good enough outcomes for the presence of outliers. So, those values are required to handle and get more improved results. Among various methods, the interquartile range (IQR) method is widely used to find different types of outliers. In IQR method, three values such as first (Q1), second (Q2), third (Q3) quartiles are considered. Then, all other values that remain outside between Q1 and Q3 are called outliers. Different instances of the dataset are arranged in ascending order and placed them into four equal sections. Since IQR expands from the first to third quartiles, then the outcomes of IQR is Q3 – Q1. Hence, all records that are under the lower limit (Q1 – 1.5 IQR) and over the upper limit (Q3 + 1.5 IQR) are called outliers. Therefore, all outliers can be detected in this way. After detecting them, they can be dropped or replaced by another suitable values. These instances affect the result of different machine learning algorithms in a particular dataset.

E. Applying Baseline Classifiers

Different types of widely used classification methods namely baseline classifiers are useful to explore various kinds of records and analyze their performance. After outlier detection and removal from primary and ANOVA F-test, CSFS, MIFS, and RFE datasets, several widely used classifiers including Gaussian Naive Bayes (GNB) [34], [35], Logistic Regression (LR) [14], [36], Random Forest (RF) [37], [38], Decision Tree (DT) [22], Extreme Gradient Boosting (XGB) [39], [11], Gradient Boosting (GB) [23], K-Nearest Neighbour (KNN) [40], AdaBoost [41], Support Vector Machine (SVM) [21], Multi-layer Perceptron (MLP) [42], and Extra Trees (ET) [43] are used to investigate PD detection dataset more precisely.

F. Evaluation Metrics

Some performance metrics such as accuracy, AUC, F-measure, Geometric mean, Sensitivity, Specificity, false positive rate, false negative rate have been used to evaluate the results of individual classifier. These metrics are expressed as a function of True Positive (TP), True Negative (TN), False Negative (FN), False Positive (FP) values.

- **Accuracy** is one of the most common evaluation metrics for classification models. It refers to how accurate a classification method is. We can express it as,

$$Accuracy = \frac{TP+TN}{TP+FN+TP+TN} \tag{1}$$

- **AUC** characterizes how well positive classes are isolated from negative classes. It can be represented with TP rate (TPR) and TN rate (TNR) by following equations:

$$\text{AUC} = \frac{TPR+TNR}{2} \tag{2}$$

- F-Measure is a harmonic mean of precision and recall.

$$F-Measure = \frac{TP}{TP+0.5(FP+FN)} \tag{3}$$

- **Geometric mean (G-mean)** is a measure of central tendency computed as the square root of specificity and sensitivity. The equation is

$$G-mean = \sqrt{\frac{TP}{TP+FN} \times \frac{TN}{TN+FP}} \tag{4}$$

- **Sensitivity** refers to the proportion of the positive events against positive predicted events. So,

$$Sensitivity = \frac{TP}{TP+FN} \tag{5}$$

- **Specificity** refers to the proportion of the negative events against predicted negative events. So,

$$Specificity = \frac{TN}{FN+FP} \tag{6}$$

- **False positive rate (Fall Out)** shows the ratio between the number of negative samples which falsely classifies as positive.

$$False\ positive\ rate = \frac{FP}{FP+TN} \tag{7}$$

- **False negative rate (Miss Rate)** shows the ratio between the number of positive samples, which falsely classified as negative.

$$False-negative\ rate = \frac{FN}{FN+TP} \tag{8}$$

IV. Experiment Result and Discussion

In this experiment, we implement different machine learning techniques such as feature selection, outlier detection and classification methods using scikit-learn library in python. From different feature subsets, we generate CFS, AVONA F-test, CSFS, MIFS, and RFS dataset as well as implemented IQR method to detect outliers. However, DT, KNN, GNB, SVM, LR, MLP, XGB, RF, ET, Adaboost, GB, and SGB has been used to investigate these subdatasets along with primary dataset. This experiment has been conducted on Google Colaboratory.

A. Performance Analysis of Classifiers for Primary Dataset

In this work, the outcomes of each classifier for primary dataset are represented at Table I. Among all classifiers, GNB provides the best findings with 82.50% accuracy, 82.50% AUC, 82.49% F-measure, 82.50% G-mean, 82.50% Sensitivity, 82.50% Specificity, and the lowest 17.50% fall out, and 17.50% miss rate. Then, LR shows the second highest results to investigate and detect PD patients. Another classifiers such as DT, KNN, SVM, MLP, XGB, RF, ET, Adaboost, and GB show good result in this work. However, MLP and SGD do not produce more improved outcomes to identify PD patients.

When we investigate various ROC curves of different classifiers, GNB provides more TPR than any other classifier (see Fig. 2). Besides, another classifiers display good TPR except MLP and SGD.

TABLE I. EXPERIMENTAL RESULTS OF DIFFERENT CLASSIFIERS FOR PRIMARY DATASET

Classifier	Accuracy	AUC	F-Measure	G-Mean	Sensitivity	Specificity	Fall Out	Miss Rate
DT	70.42	70.42	70.33	70.42	70.42	70.42	29.58	29.58
KNN	72.08	72.08	72.08	72.08	72.08	72.08	27.92	27.92
GNB	**82.50**	**82.50**	**82.50**	**82.50**	**82.50**	**82.50**	**17.50**	**17.50**
SVM	72.92	72.92	72.86	72.92	72.92	72.92	27.08	27.08
LR	77.08	77.08	77.08	77.08	77.08	77.08	22.92	22.92
MLP	57.08	57.08	54.70	57.08	57.08	57.08	42.92	42.92
XGB	74.58	74.58	74.55	74.58	74.58	74.58	25.42	25.42
RF	75.83	75.83	75.73	75.83	75.83	75.83	24.17	24.17
ET	77.08	77.08	76.91	77.08	77.08	77.08	22.92	22.92
AdaBoost	70.00	70.00	70.00	70.00	70.00	70.00	30.00	30.00
GB	70.83	70.83	70.76	70.83	70.83	70.83	29.17	29.17
SGD	51.67	51.67	51.33	51.67	51.67	51.67	48.33	48.33

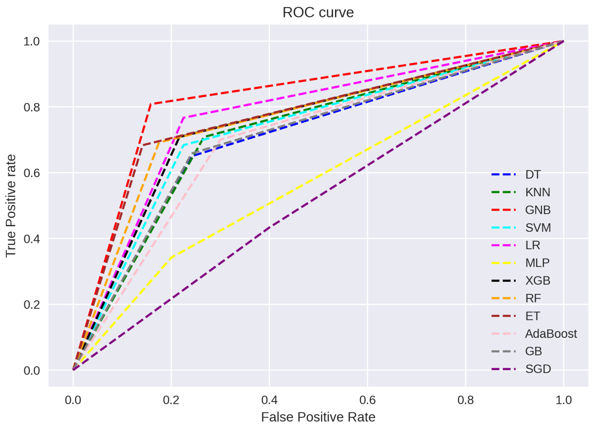

Fig. 2. ROC Curves of Individual Classifiers for Primary Dataset.

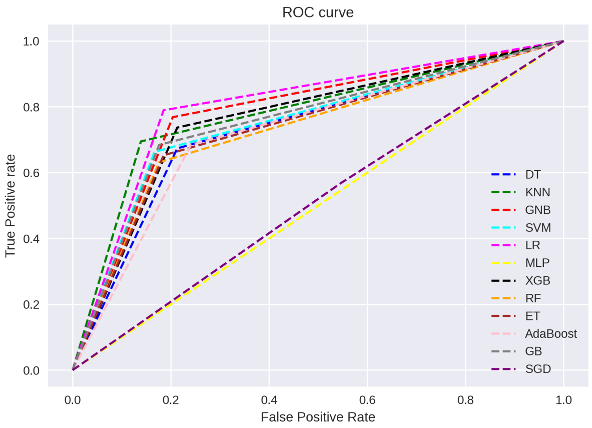

Fig. 3. ROC Curves of Individual Classifiers for CFS Dataset.

B. Performance Analysis of Classifiers for CFS Dataset

According to the outcomes at Table II, LR obtains the best 80.30% accuracy, 80.21% AUC, 80.30% f-measure, 80.21% g-mean, 80.30% sensitivity, 80.13% specificity where it shows 19.87% fall out and 19,70% miss rate. However, it does not exceed the highest of GNB for primary dataset. The results of several classifiers such as DT, KNN, SVM, XGB, Adaboost, and GB are improved for CFS than primary dataset. Instead, GNB, MLP, RF, and ET are slightly decreased than primary dataset in this work.

After observing ROC curves of each classifier, LR also shows more TPR than other classifiers (see Fig. 3). However, MLP and SGD do not provide good TPR like most of the classifiers in this work.

C. Performance Analysis of Classifiers for ANOVA F-test Dataset

In the classification result of Table III, GNB obtained the best outcomes for ANOVA F-test dataset and does not give improved results compared to primary dataset (81.42% accuracy, 81.41% AUC, 81.42% F-measure, 81.41% G-mean, 81.42% Sensitivity, 81.40% Specificity, 18.60% fall out, 18.58% miss out). Also, the degradation of results are noticed for KNN, SVM, MLP, RF, ET, and SGD. However, we noticed a performance boost for DT, LR, XGB, AdaBoost, and GB respectively.

Then, when we consider ROC curves of different classifier at Fig. 4, GNB shows the highest TPR to detect PD more precisely. Besides, LR, DT, KNN, SVM, XGB, RF, ET, Adaboost, and GB also represent good outcomes in this work.

D. Performance Analysis of Classifiers for CSFS Dataset

Then, GNB gives the best performance (80% accuracy, 79.74% AUC, 79.87% F-measure, 79.74% G-mean, 80% Sensitivity, 79.48% Specificity, 20.52% fall out, 20% miss rate) whereas it does not exceed the outcomes for primary dataset (see Table IV). Also, many classifiers like KNN, SVM, MLP, XGB, RF, ET, AdaBoost, and GB are not generated good results where DT, LR, and SGD show improved results than primary dataset.

TABLE II. Experimental Results of Different Classifiers for CFS Dataset

Classifier	Accuracy	AUC	F-Measure	G-Mean	Sensitivity	Specificity	Fall Out	Miss Rate
DT	73.40	73.04	73.29	73.04	73.40	72.67	27.33	26.60
KNN	78.33	77.79	78.12	77.79	78.33	77.26	22.74	21.67
GNB	78.33	78.24	78.33	78.24	78.33	78.15	21.85	21.67
SVM	75.37	74.82	75.14	74.82	75.37	74.28	25.72	24.63
LR	**80.30**	**80.21**	**80.30**	**80.21**	**80.30**	**80.13**	**19.87**	**19.70**
MLP	53.20	50.00	36.95	49.90	53.20	46.80	53.20	46.80
XGB	76.35	76.19	76.34	76.19	76.35	76.03	23.97	23.65
RF	73.40	72.78	73.09	72.78	73.40	72.17	27.83	26.60
ET	73.89	73.37	73.67	73.37	73.89	72.85	27.15	26.11
AdaBoost	72.41	72.17	72.37	72.17	72.41	71.93	28.07	27.59
GB	75.86	75.41	75.71	75.41	75.86	74.97	25.03	24.14
SGD	50.74	51.11	50.63	51.10	50.74	51.47	48.53	49.26

TABLE III. Experimental Results of Different Classifiers for ANOVA F-test Dataset

Classifier	Accuracy	AUC	F-Measure	G-Mean	Sensitivity	Specificity	Fall Out	Miss Rate
DT	73.89	73.86	73.88	73.86	73.89	73.83	26.17	26.11
KNN	68.58	68.47	68.45	68.47	68.58	68.36	31.64	31.42
GNB	**81.42**	**81.41**	**81.42**	**81.41**	**81.42**	**81.40**	**18.60**	**18.58**
SVM	68.58	68.47	68.45	68.47	68.58	68.36	31.64	31.42
LR	79.65	79.64	79.65	79.64	79.65	79.63	20.37	20.35
MLP	50.88	50.00	34.32	49.99	50.88	49.12	50.88	49.12
XGB	76.55	76.52	76.54	76.52	76.55	76.49	23.51	23.45
RF	75.22	75.12	75.13	75.12	75.22	75.02	24.98	24.78
ET	71.68	71.61	71.63	71.61	71.68	71.54	28.46	28.32
AdaBoost	73.45	73.43	73.45	73.43	73.45	73.40	26.60	26.55
GB	72.12	72.14	72.13	72.14	72.12	72.15	27.85	27.88
SGD	50.00	49.48	45.16	49.47	50.00	48.95	51.05	50.00

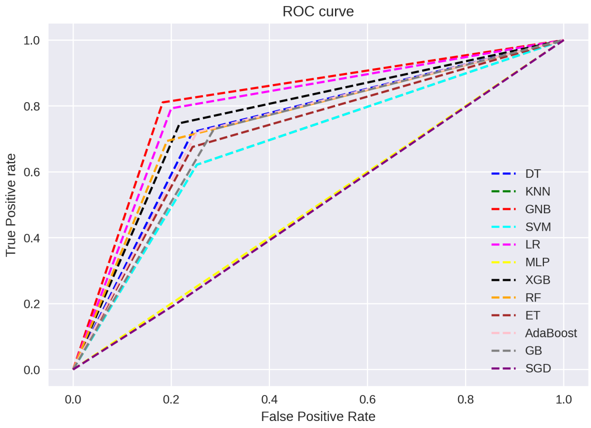

Fig. 4. ROC Curves of Individual Classifiers for ANOVA F-test Dataset.

When the ROC curves of different classifiers are observed (see Fig. 5), the curves of GNB and LR are very close to each other, but GNB is the best classifier to represent this curve. Again, MLP and SGD show its low TPR for CSFS dataset analysis.

E. Performance Analysis of Classifiers for MIFS Dataset

In this case, the outcomes of GNB and LR are very close to each other (see Table V). But, GNB shows slightly improved result than LR (79.29% accuracy, 79.28% AUC, 79.29% F-measure, 0.7928 G-mean, 79.29% Sensitivity, 79.28% Specificity, 20.71% fall out, and 20.7% miss rate). But it is not exceed GNB result for primary dataset. However, some classifiers like KNN, SVM, MLP, RF, ET, and SGD provide worsen results in MIFS dataset. However, the results of DT, LR, XGB, Adaboost, and GB are given a few improved result for MIFS than primary dataset.

However, the ROC curve of GNB and LR are almost same for MIFS dataset (see Fig. 6). Another classifiers also display good ROC curve except MLP and SGD.

TABLE IV. EXPERIMENTAL RESULTS OF DIFFERENT CLASSIFIERS FOR CSFS DATASET

Classifier	Accuracy	AUC	F-Measure	G-Mean	Sensitivity	Specificity	Fall Out	Miss Rate
DT	70.45	70.33	70.41	70.33	70.45	70.21	29.79	29.55
KNN	70.91	70.74	70.83	70.74	70.91	70.57	29.43	29.09
GNB	**80.00**	**79.74**	**79.87**	**79.74**	**80.00**	**79.48**	**20.52**	**20.00**
SVM	67.73	67.50	67.59	67.50	67.73	67.28	32.72	32.27
LR	78.18	78.02	78.12	78.02	78.18	77.86	22.14	21.82
MLP	51.36	49.56	35.17	49.53	51.36	47.76	52.24	48.64
XGB	71.36	71.14	71.24	71.14	71.36	70.92	29.08	28.64
RF	72.73	72.46	72.55	72.46	72.73	72.19	27.81	27.27
ET	71.82	71.52	71.59	71.52	71.82	71.21	28.79	28.18
AdaBoost	68.18	68.24	68.19	68.24	68.18	68.30	31.70	31.82
GB	67.27	67.16	67.24	67.16	67.27	67.05	32.95	32.73
SGD	55.00	53.70	48.25	53.68	55.00	52.40	47.60	45.00

TABLE V. EXPERIMENTAL RESULTS OF DIFFERENT CLASSIFIERS FOR MIFS DATASET

Classifier	Accuracy	AUC	F-Measure	G-Mean	Sensitivity	Specificity	Fall Out	Miss Rate
DT	74.01	73.99	73.94	73.99	74.01	73.97	26.03	25.99
KNN	70.04	70.03	70.00	70.03	70.04	70.01	29.99	29.96
GNB	**79.30**	**79.29**	**79.29**	**79.29**	**79.30**	**79.28**	**20.72**	**20.70**
SVM	69.16	69.16	69.16	69.16	69.16	69.15	30.85	30.84
LR	**79.30**	**79.29**	79.28	**79.29**	**79.30**	79.27	20.73	**20.70**
MLP	50.22	50.00	33.58	50.00	50.22	49.78	50.22	49.78
XGB	74.89	74.87	74.85	74.87	74.89	74.85	25.15	25.11
RF	72.25	72.22	72.15	72.22	72.25	72.20	27.80	27.75
ET	75.33	75.31	75.30	75.31	75.33	75.30	24.70	24.67
AdaBoost	72.69	72.67	72.65	72.67	72.69	72.66	27.34	27.31
GB	72.25	72.24	72.25	72.24	72.25	72.24	27.76	27.75
SGD	50.22	50.26	49.73	50.26	50.22	50.31	49.69	49.78

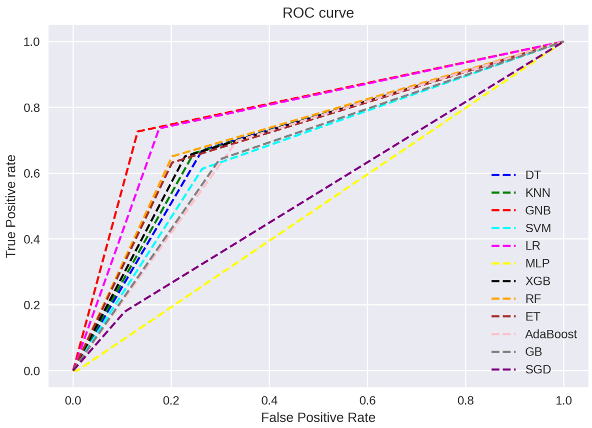

Fig. 5. ROC Curves of Individual Classifiers for CSFS Dataset.

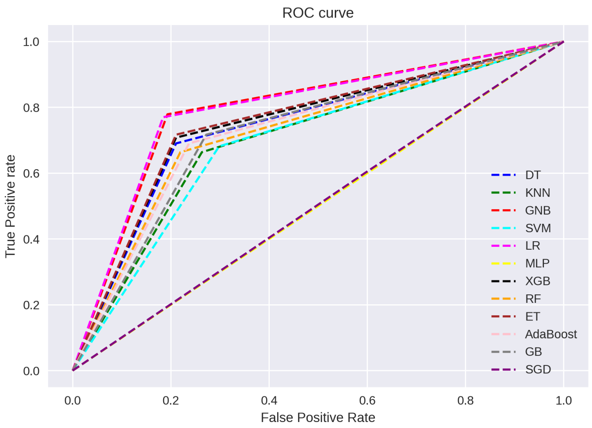

Fig. 6. ROC Curves of Individual Classifiers for MIFS Dataset.

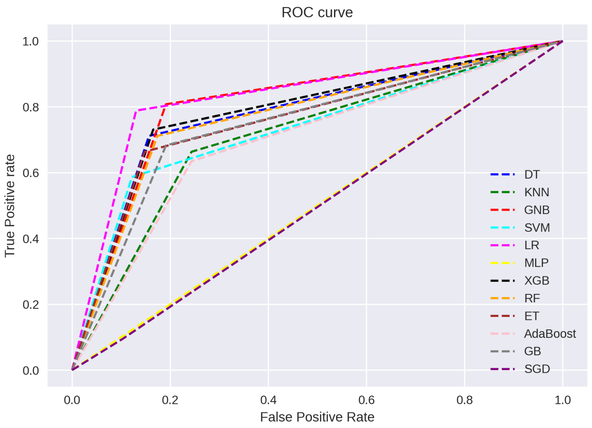

Fig. 7. ROC Curves of Individual Classifiers for RFE Dataset.

F. Performance Analysis of Classifiers for RFE Dataset

In this case, Table VI shows individual performance for RFE dataset where LR shows the highest result with 83.11% accuracy, 82.90% AUC, 83.06% F-measure, 82.90% G-mean, 83.11% Sensitivity, 82.70% Specificity, 17.30% fall out, and 16.89% miss rate. Therefore, it outperforms the best performance of GNB for primary dataset. A few improved results are found for some classifier excluding KNN, GNB, MLP, ET, AdaBoost, and SGB for RFE dataset.

Also, LR shows the best ROC curve whose represent more TPR than any other classifier for RFE dataset (see Fig. 7).

As we observe the performance measures and ROC curves of different classifier, LR determine the best outcomes for RFE dataset. But, these results are not found more stable in various cases. After observing the results of primary and its generated subdatasets, different classifiers give better outcomes and feature reduction methods are shown effective findings to detect PD patients. Also, we scrutinize the average results of different classifier which represents at Table VII. In this case, GNB displays the best average outcomes among all classifiers. Likewise, LR provides the second highest average outcomes in this analysis. Then, RF, XGB, ET, DT, KNN, GB, and AdaBoost give well average results like previous observations in the primary and its sub datasets. MLP and SGD do not represent good average outcomes in this work.

This proposed framework is integrated more feature selection and classification method than other existing works [14], [17], [20], [8], [44]. To evaluate its results, we consider various kinds of evaluation metrics where different previous works [21], [2], [23] has not maintained such types of evaluation. Along with best feature selection and classification methods, this framework also explores the most stable classifier which can provide better outcome in any types of transformation and experimental settings.

V. Conclusion and Future Work

This research has identified a reliable technique for feature selection of PD dataset with more simplicity, less running time, and cost-effectiveness. First, we explore insignificant features using different methods, remove them and generate sub datasets. However, the IQR method has been applied to detect outliers and prune them. Then, a lot of classifiers are used to investigate different types of PD datasets and compared them with primary dataset. In this case, LR shows the highest outcomes for RFE-based method. Besides, GNB is the most stable method to investigate Parkinson acoustic instances. This method can be potentially applied to similar types of datasets to obtain better solutions, distinguish between normal and sick people, and lessen diagnosis costs. Some feature selection and classification methods are provided random outcomes due to some infrastructural settings. In future, we would like to work on different limitations and gathered more widely used technologies to provide more satisfactory outcomes for detecting PD.

Acknowledgment

The authors would like to thank Research Cell of Noakhali Science and Technology University, Noakhali-3814, Bangladesh for supporting with the fund.

References

[1] J. Jankovic, "Parkinson's disease: clinical features and diagnosis," *Journal of neurology, neurosurgery & psychiatry*, vol. 79, no. 4, pp. 368–376, 2008.

[2] G. Pahuja and T. Nagabhushan, "A comparative study of existing machine learning approaches for parkinson's disease detection," *IETE Journal of Research*, vol. 67, no. 1, pp. 4–14, 2021.

[3] N. Singh, V. Pillay, and Y. E. Choonara, "Advances in the treatment of parkinson's disease," *Progress in neurobiology*, vol. 81, no. 1, pp. 29–44, 2007.

[4] E. S. Levy, G. Moya-Galé, Y. H. M. Chang, K. Freeman, K. Forrest, M. F. Brin, and L. A. Ramig, "The effects of intensive speech treatment on intelligibility in parkinson's disease: a randomised controlled trial," *EClinicalMedicine*, vol. 24, p. 100429, 2020.

[5] M. Nilashi, O. Ibrahim, H. Ahmadi, L. Shahmoradi, and M. Farahmand, "A hybrid intelligent system for the prediction of parkinson's disease progression using machine learning techniques," *Biocybernetics and Biomedical Engineering*, vol. 38, no. 1, pp. 1–15, 2018.

[6] Z. K. Senturk, "Early diagnosis of parkinson's disease using machine learning algorithms," *Medical hypotheses*, vol. 138, p. 109603, 2020.

[7] H. N. Pham, T. T. Do, K. Y. J. Chan, G. Sen, A. Y. Han, P. Lim, T. S. L. Cheng, Q. H. Nguyen, B. P. Nguyen, and M. C. Chua, "Multimodal detection of parkinson disease based on vocal and improved spiral test," in *2019 International Conference on System Science and Engineering (ICSSE)*. IEEE, 2019, pp. 279–284.

[8] K. N. R. Challa, V. S. Pagolu, G. Panda, and B. Majhi, "An improved approach for prediction of parkinson's disease using machine learning techniques," in *2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES)*. IEEE, 2016, pp. 1446–1451.

[9] M. S. Satu, S. Roy, F. Akhter, and M. Whaiduzzaman, "IoLT: an IOT based collaborative blended learning platform in higher education," in *2018 International Conference on Innovation in Engineering and technology (ICIET)*. IEEE, 2018, pp. 1–6.

[10] M. S. Satu, M. I. Khan, M. R. Rahman, K. C. Howlader, S. Roy, S. S. Roy, J. M. Quinn, and M. A. Moni, "Diseasome and comorbidities complexities of sars-cov-2 infection with common malignant diseases," *Briefings in Bioinformatics*, vol. 22, no. 2, pp. 1415–1429, 2021.

[11] M. S. Satu, M. I. Khan, M. Mahmud, S. Uddin, M. A. Summers, J. M. Quinn, and M. A. Moni, "Tclustvid: a novel machine learning classification model to investigate topics and sentiment in covid-19 tweets," *Knowledge-Based Systems*, vol. 226, p. 107126, 2021.

[12] K. Ahammed, M. S. Satu, M. I. Khan, and M. Whaiduzzaman, "Predicting infectious state of hepatitis c virus affected patient's applying machine learning methods," in *2020 IEEE Region 10 Symposium (TENSYMP)*. IEEE, 2020, pp. 1371–1374.

TABLE VI. EXPERIMENTAL RESULTS OF DIFFERENT CLASSIFIERS FOR RFE DATASET

Classifier	Accuracy	AUC	F-Measure	G-Mean	Sensitivity	Specificity	Fall Out	Miss Rate
DT	78.08	77.75	77.95	77.75	78.08	77.42	22.58	21.92
KNN	71.23	71.00	71.16	71.00	71.23	70.77	29.23	28.77
GNB	80.82	80.82	80.83	80.82	80.82	80.82	19.18	19.18
SVM	73.97	73.24	73.32	73.24	73.97	72.51	27.49	26.03
LR	**83.11**	**82.90**	**83.06**	**82.90**	**83.11**	**82.70**	**17.30**	**16.89**
MLP	52.51	50.00	36.16	49.94	52.51	47.49	52.51	47.49
XGB	78.54	78.28	78.46	78.28	78.54	78.02	21.98	21.46
RF	77.17	76.88	77.07	76.88	77.17	76.59	23.41	22.83
ET	76.26	75.78	75.99	75.78	76.26	75.31	24.69	23.74
AdaBoost	69.86	69.56	69.73	69.56	69.86	69.25	30.75	30.14
GB	74.89	74.57	74.76	74.57	74.89	74.25	25.75	25.11
SGD	51.60	49.59	42.21	49.55	51.60	47.58	52.42	48.40

TABLE VII. AVERAGE CLASSIFICATION RESULTS OF INDIVIDUAL CLASSIFIERS

Classifier	Accuracy	AUC	F-Measure	G-Mean	Sensitivity	Specificity	Fall Out	Miss Rate
DT	73.38	73.23	73.30	73.23	73.38	73.09	26.91	26.62
KNN	71.86	71.69	71.77	71.69	71.86	71.51	28.49	28.14
GNB	**80.39**	**80.33**	**80.37**	**80.33**	**80.39**	**80.27**	**19.73**	**19.61**
SVM	71.29	71.02	71.09	71.02	71.29	70.75	29.25	28.71
LR	79.60	79.52	79.58	79.52	79.60	79.45	20.55	20.40
MLP	52.54	51.11	38.48	51.07	52.54	49.67	50.33	47.46
XGB	75.38	75.26	75.33	75.26	75.38	75.15	24.85	24.62
RF	74.43	74.22	74.29	74.22	74.43	74.00	26.00	25.57
ET	74.34	74.11	74.18	74.11	74.34	73.88	26.12	25.66
AdaBoost	71.10	71.01	71.06	71.01	71.10	70.92	29.08	28.90
GB	72.20	72.06	72.14	72.06	72.20	71.92	28.08	27.80
SGD	51.54	50.97	47.88	50.96	51.54	50.40	49.60	48.46

[13] T. Akter, M. H. Ali, M. Khan, M. Satu, M. Uddin, S. A. Alyami, S. Ali, A. Azad, M. A. Moni *et al.*, “Improved transfer-learning-based facial recognition framework to detect autistic children at an early stage,” *Brain Sciences*, vol. 11, no. 6, p. 734, 2021.

[14] R. Das, “A comparison of multiple classification methods for diagnosis of parkinson disease,” *Expert Systems with Applications*, vol. 37, no. 2, pp. 1568–1572, 2010.

[15] A. Tsanas, M. A. Little, C. Fox, and L. O. Ramig, “Objective automatic assessment of rehabilitative speech treatment in parkinson’s disease,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 22, no. 1, pp. 181–190, 2013.

[16] N. Shamli, B. Sathiyabhama *et al.*, “Parkinson’s brain disease prediction using big data analytics,” *International Journal of Information Technology and Computer Science (IJITCS)*, vol. 8, no. 6, p. 73, 2016.

[17] T. Tong, C. Ledig, R. Guerrero, A. Schuh, J. Koikkalainen, A. Tolonen, H. Rhodius, F. Barkhof, B. Tijms, A. W. Lemstra *et al.*, “Five-class differential diagnostics of neurodegenerative diseases using random undersampling boosting,” *NeuroImage: Clinical*, vol. 15, pp. 613–624, 2017.

[18] W. Lu, Z. Li, and J. Chu, “A novel computer-aided diagnosis system for breast mri based on feature selection and ensemble learning,” *Computers in biology and medicine*, vol. 83, pp. 157–165, 2017.

[19] R. Mathur, V. Pathak, and D. Bandil, “Parkinson disease prediction using machine learning algorithm,” in *Emerging Trends in Expert Applications and Security*. Springer, 2019, pp. 357–363.

[20] J. S. Almeida, P. P. Rebouças Filho, T. Carneiro, W. Wei, R. Damaševičius, R. Maskeliūnas, and V. H. C. de Albuquerque, “Detecting parkinson’s disease with sustained phonation and speech signals using machine learning techniques,” *Pattern Recognition Letters*, vol. 125, pp. 55–62, 2019.

[21] S. Lahmiri and A. Shmuel, “Detection of parkinson’s disease based on voice patterns ranking and optimized support vector machine,” *Biomedical Signal Processing and Control*, vol. 49, pp. 427–433, 2019.

[22] S. A. Mostafa, A. Mustapha, M. A. Mohammed, R. I. Hamed, N. Arunkumar, M. K. Abd Ghani, M. M. Jaber, and S. H. Khaleefah, “Examining multiple feature evaluation and classification methods for improving the diagnosis of parkinson’s disease,” *Cognitive Systems Research*, vol. 54, pp. 90–99, 2019.

[23] I. Karabayir, S. M. Goldman, S. Pappu, and O. Akbilgic, “Gradient boosting for parkinson’s disease diagnosis from voice recordings,” *BMC Medical Informatics and Decision Making*, vol. 20, no. 1, pp. 1–7, 2020.

[24] R. Lamba, T. Gulati, H. F. Alharbi, and A. Jain, “A hybrid system for parkinson’s disease diagnosis using machine learning techniques,” *International Journal of Speech Technology*, pp. 1–11, 2021.

[25] M. Pramanik, R. Pradhan, P. Nandy, A. K. Bhoi, and P. Barsocchi, “Machine learning methods with decision forests for parkinson’s detection,” *Applied Sciences*, vol. 11, no. 2, p. 581, 2021.

[26] L. Naranjo, C. J. Perez, Y. Campos-Roca, and J. Martin, “Addressing voice recording replications for parkinson’s disease detection,” *Expert Systems with Applications*, vol. 46, pp. 286–292, 2016.

[27] T. Akter, M. I. Khan, M. H. Ali, M. S. Satu, M. J. Uddin, and M. A. Moni, "Improved machine learning based classification model for early autism detection," in *2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST)*. IEEE, 2021, pp. 742–747.

[28] M. S. Siraj, M. A. A. Faisal, O. Shahid, F. F. Abir, T. Hossain, S. Inoue, and M. A. R. Ahad, "Upic: user and position independent classical approach for locomotion and transportation modes recognition," in *Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers*, 2020, pp. 340–345.

[29] I. S. Thaseen and C. A. Kumar, "Intrusion detection model using fusion of chi-square feature selection and multi class svm," *Journal of King Saud University-Computer and Information Sciences*, vol. 29, no. 4, pp. 462–472, 2017.

[30] J. R. Vergara and P. A. Estévez, "A review of feature selection methods based on mutual information," *Neural computing and applications*, vol. 24, no. 1, pp. 175–186, 2014.

[31] H. Nkiama, S. Z. M. Said, and M. Saidu, "A subset feature elimination mechanism for intrusion detection system," *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 4, pp. 148–157, 2016.

[32] M. S. Satu, K. Howlader, M. P. Hosen, N. Chowdhury, and M. A. Moni, "Identifying the stability of couple relationship applying different machine learning techniques," in *2020 11th International Conference on Electrical and Computer Engineering (ICECE)*. IEEE, 2020, pp. 246–249.

[33] M. S. Satu, S. T. Atik, and M. A. Moni, "A novel hybrid machine learning model to predict diabetes mellitus," in *Proceedings of International Joint Conference on Computational Intelligence*. Springer, 2020, pp. 453–465.

[34] A. C. A. de Araújo, E. G. d. R. Santos, K. S. G. de Sá, V. K. T. Furtado, F. A. Santos, R. C. de Lima, L. V. Krejcová, B. L. Santos-Lobato, G. H. L. Pinto, A. d. S. Cabral *et al.*, "Hand resting tremor assessment of healthy and patients with parkinson's disease: An exploratory machine learning study," *Frontiers in bioengineering and biotechnology*, vol. 8, p. 778, 2020.

[35] Nurjahan, M. A. T. Rony, M. S. Satu, M. Whaiduzzaman *et al.*, "Mining significant features of diabetes through employing various classification methods," in *2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD)*. IEEE, 2021, pp. 240–244.

[36] M. S. Satu, K. Mizan, S. A. Jerin, M. Whaiduzzaman, A. Barros, K. Ahmed, M. A. Moni *et al.*, "Covid-hero: Machine learning based covid-19 awareness enhancement mobile game for children," in *International Conference on Applied Intelligence and Informatics*. Springer, 2021, pp. 321–335.

[37] A. K. Tiwari, "Machine learning based approaches for prediction of parkinson's disease," *Mach Learn Appl*, vol. 3, no. 2, pp. 33–39, 2016.

[38] S. Rahman, M. I. Khan, M. S. Satu, and M. Z. Abedin, "Risk prediction with machine learning in cesarean section: Optimizing healthcare operational decisions," in *Signal Processing Techniques for Computational Health Informatics*. Springer, 2021, pp. 293–314.

[39] H. C. Tunc, C. O. Sakar, H. Apaydin, G. Serbes, A. Gunduz, M. Tutuncu, and F. Gurgen, "Estimation of parkinson's disease severity using speech features and extreme gradient boosting," *Medical & Biological Engineering & Computing*, vol. 58, no. 11, pp. 2757–2773, 2020.

[40] L. Chen and M. S. Kamel, "Msebag: a dynamic classifier ensemble generation based on 'minimum-sufficient ensemble'and bagging," *International Journal of Systems Science*, vol. 47, no. 2, pp. 406–419, 2016.

[41] V. K. Gudipati, O. R. Barman, M. Gaffoor, A. Abuzneid *et al.*, "Efficient facial expression recognition using adaboost and haar cascade classifiers," in *2016 Annual Connecticut Conference on Industrial Electronics, Technology & Automation (CT-IETA)*. IEEE, 2016, pp. 1–4.

[42] S. Wan, Y. Liang, Y. Zhang, and M. Guizani, "Deep multi-layer perceptron classifier for behavior analysis to estimate parkinson's disease severity using smartphones," *IEEE Access*, vol. 6, pp. 36 825–36 833, 2018.

[43] E. Celik and S. I. Omurca, "Improving parkinson's disease diagnosis with machine learning methods," in *2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT)*. IEEE, 2019, pp. 1–4.

[44] R. Domingues, M. Filippone, P. Michiardi, and J. Zouaoui, "A comparative evaluation of outlier detection algorithms: Experiments and analyses," *Pattern Recognition*, vol. 74, pp. 406–421, 2018.