IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 4, 2019, accepted July 15, 2019, date of publication July 30, 2019, date of current version September 3, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2932037

Reliable Parkinson's Disease Detection by Analyzing Handwritten Drawings: Construction of an Unbiased Cascaded Learning System Based on Feature Selection and Adaptive Boosting Model

LIAQAT ALI[1], CE ZHU[1], (Fellow, IEEE), NOORBAKHSH AMIRI GOLILARZ[2], ASHIR JAVEED[3], MINGYI ZHOU[1], AND YIPENG LIU[1], (Senior Member, IEEE)

[1]School of Information and Communication Engineering, University of Electronic Science and Technology of China (UESTC), Chengdu 611731, China
[2]School of Computer Science and Engineering, University of Electronic Science and Technology of China (UESTC), Chengdu 611731, China
[3]School of Information and Software Engineering, University of Electronic Science and Technology of China (UESTC), Chengdu 611731, China

Corresponding authors: Liaqat Ali (engr_liaqat183@yahoo.com) and Ce Zhu (eczhu@uestc.edu.cn)

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61602091 and Grant 61571102, and in part by the Sichuan Science and Technology Program under Grant 2019YFH0008 and Grant 2018JY0035.

ABSTRACT Parkinson's disease (PD) is the second most common neurodegenerative disease of central nervous system (CNS). Till now, there is no definitive clinical examination that can diagnose a PD patient. However, it has been reported that PD patients face deterioration in handwriting. Hence, different computer vision and machine learning researchers have proposed micrography and computer vision based methods. But, these methods possess two main problems. The first problem is biasedness in models caused by imbalanced data *i.e.* machine learning models show good performance on majority class but poor performance on minority class. Unfortunately, previous studies neither discussed this problem nor took any measures to avoid it. In order to highlight the biasedness in the constructed models and practically demonstrate it, we develop four different machine learning models. To alleviate the problem of biasedness, we propose to use random undersampling method to balance the training process. The second problem is low rate of classification accuracy which has limited clinical significance. To improve the PD detection accuracy, we propose a cascaded learning system that cascades a Chi2 model with adaptive boosting (Adaboost) model. The Chi2 model ranks and selects a subset of relevant features from the feature space while Adaboost model is used to predict PD based on the subset of features. Experimental results confirm that the proposed cascaded system shows better performance than other six similar cascaded systems that used six different state of the art machine learning models. Moreover, it was also observed that the proposed cascaded system improves the strength of conventional Adaboost model by 3.3% and reduces its complexity. Additionally, the cascaded system achieved classification accuracy of 76.44%, sensitivity of 70.94% and specificity of 81.94%.

INDEX TERMS Balanced accuracy, machine learning, oversampling, Parkinson's disease, undersampling.

I. INTRODUCTION

Parkinson's disease (PD) is reported to be the second most common neurological syndrome of the central nervous system after Alzheimer's disease (AD) [1]. PD targets elder people mostly having age of 60 years or above [2]. The most common symptoms observed in PD patients include bradykinesia (slowness of movement), dysphonia (voice impairments), rigidity, tremor, and poor balance [3]–[7]. Till now, the recognition of PD is a clinically challenging task [8]–[10]. However, it is known that PD patients face deterioration in handwriting. Hence, different computer vision and machine learning researchers have proposed micrography and vision based methods to automatically detect PD using handwritten exams.

Drotar *et al.* in [11] pointed out that in-air movements during handwriting have a major impact on the PD detection accuracy. Rosenblum *et al.* pointed out that PD patients can be discriminated from healthy subjects using handwriting exams [12]. They conducted a study on 20 PD patients and

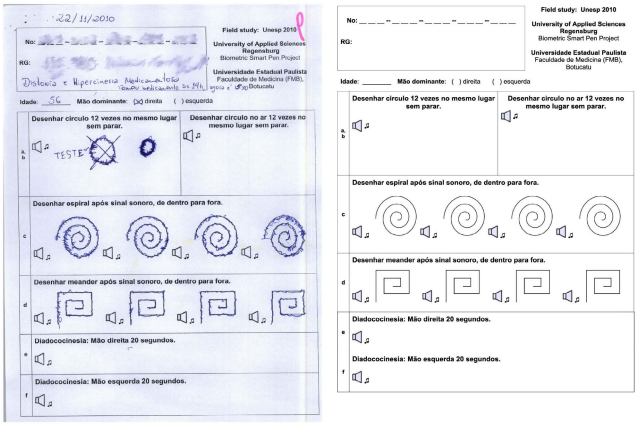
The associate editor coordinating the review of this article and approving it for publication was Larbi Boubchir.

20 healthy subjects. Each subject was asked to write his/her name and address on a piece of paper attached to a digital table. They used mean pressure and velocity related parameters and achieved detection accuracy of 97.5% *i.e.* they achieved 100% of specificity and 95% of sensitivity. The key issue with these methods and datasets is their limited size. Hence, the results have limited significance. Thus, Drotar *et al.* in [13] collected data from 37 PD patients and 38 control subjects by performing eight different hand writing tasks. They developed three different machine learning models namely k-nearest neighbors (KNN), Adaboost ensemble model and support vector machine (SVM) and achieved classification accuracy of 81.3%.

Recently, Pereira *et al.* developed computer vision and machine learning based methods to contribute in the process of PD detection [14]–[18]. Pereira *et al.* collected data from 55 individuals consisting of 37 PD patients and 18 healthy subjects [14]. The dataset contained only spiral drawings. They utilized optimum path forest (OPF), SVM and Naive Bayes (NB) supervised models for discriminating the hand written drawings of PD patients from that of healthy subjects. They achieved best accuracy of 78.9% with NB model. In their future studies, they developed another dataset which was collected from 18 healthy subjects and 74 PD patients [15]. The dataset was named HandPD which is so far the largest publicly available handwritten dataset having 736 samples. The HandPD dataset contains spiral and meander drawings. For the HandPD dataset, features from the hand written drawings were extracted using computer vision based methods and for classification same three models *i.e.* OPF, SVM and NB were utilized. They achieved classification accuracy of 67% for HandPD dataset. There are two main problems in studies conducted on the HandPD dataset. That is biasedness in models and low rate of PD detection accuracy.

In this paper, we consider the two problems in PD detection based on HandPD data *i.e.* the problem of biasedness in the constructed models and the low rate of PD detection accuracy. To demonstrate the problem of biasedness in models, we develop and train four different machine learning models namely Linear Discriminant Analysis (LDA), K Nearest Neighbors (KNN), Gaussian Naive Bayes (GNB) and Decision Tree (DT). To solve this problem, we utilize random undersampling method. As discussed above, the second problem in PD detection based on HandPD data is low rate of PD detection accuracy. To alleviate this problem, we propose a cascaded learning system that cascades Chi2 model with adaptive boosting (Adaboost) model. The Chi2 model is used to rank and select relevant subset of features while Adaboost is used for classification purposes. Experimental results evidently show that the proposed methods help alleviate both the problems to some extent.

The rest of the paper is organized as follows: In section II, details about the dataset are given, section III discusses problems and proposed solutions. Section IV is about validation scheme and evaluation methods. Section V is about experiments and discussion and the last section concludes the study.



(a) (b)

FIGURE 1. Sample of a form filled by a 56 years old PD patient (a) and sample of an empty form (b).

II. DESCRIPTION OF THE HANDPD DATASET AND FEATURE EXTRACTION

In this study, we adopted the HandPD dataset that is available online [19]. The dataset was collected from 92 subjects at the Faculty of Medicine of Botucatu, Sao Paulo State University, Brazil. The dataset was collected from two groups of subjects: (i) the first group contains 74 subjects which are PD patients; (ii) the second group contains 18 subjects which are healthy individuals. The first group is subdivided into 59 male and 15 female subjects while the second group is subdivided into 6 male and 12 female. Hence, the 19.56% of the whole dataset is composed of healthy subjects and 80.44% of the whole dataset is PD patients. Furthermore, the control group consists of 16 right-handed and 2 left-handed subjects. On the other hand, the PD patients group consists of 69 right-handed and 5 left-handed subjects.

During the data collection process each subject was asked to perform 6 different tasks which are shown in the FIGURE 1a-f. The figure is a form that was filled by a 56 years old PD patient. Among the six different tasks, the HandPD dataset records only two tasks *i.e.* spiral drawings and meander drawings. From each subject, 4 spirals and 4 meanders were collected. Thus, the dataset contains $92 \times 8 = 736$ drawings. Among these drawings, half *i.e.* 368 are spirals and half are meanders.

After the data collection, feature extraction process was performed. Each image (filled form) was segmented into 8 parts *i.e.* 4 meanders numbered from 1 to 4 and 4 spirals which were numbered from 5 to 8. From each of the 8 drawings numbered from 1 to 8, nine numeric features were extracted. The feature extraction process was divided into two steps. In first step, an automated method was developed to automatically separate the hand written trace (HT) from the exam template (ET) for the drawings *i.e.* spirals and meanders as shown in FIGURE 2.

In the second step, 9 statistical features were evaluated by comparing ET and HT *i.e.* by calculating the amount of difference between them. The difference between the two images

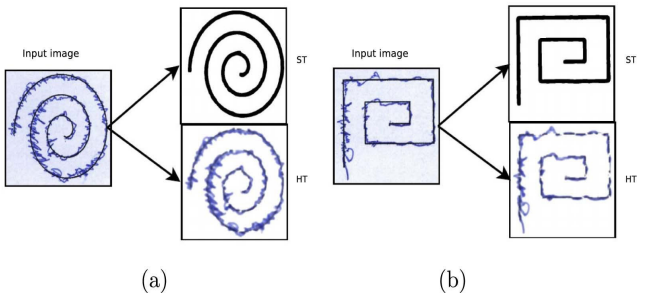


FIGURE 2. (a): Spiral image and its corresponding HT and ET, (b): meander image and its corresponding HT and ET.

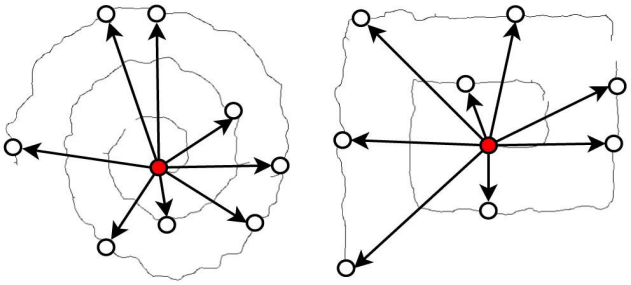


FIGURE 3. Different random points taken on spiral and meander images. Each arrow or vector is originating from the center point of spiral or meander and end up at the randomly chosen point.

was evaluated by considering a number of points sampled at same positions in HT and ET images. Before discussing the set of extracted features, it is important to discuss the radius of spiral or meander point. It is the distance of the straight line that connects the center of a spiral or meander (shown as red point in FIGURE 3) to a sampled point under consideration (white points in FIGURE 3 are the sampled points). For more details on the feature extraction process, readers are referred to [14], [15]. The extracted features are briefly discussed as follows:

C_1: The first feature denotes the Root Mean Square (RMS) of the difference between HT and ET radius which is calculated according to the formulation as follows:

$$RMS = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(r_{HT}^{i} - r_{ET}^{i})^2} \tag{1}$$

where n = the number of sampled points, r_{HT}^{i} = HT radius of i-th sample point and r_{ET}^{i} = ET radius considering the i-th sample point.

C_2: The second feature is the maximum difference between ET and HT radius which is calculated according to the formulation as follows:

$$d_{max} = \arg\max_{i}\{|r_{HT}^{i} - r_{ET}^{i}|\} \tag{2}$$

C_3: The third feature is the minimum difference between ET and HT radius, it is denoted by d_{min} in the following equation.

$$d_{min} = \arg\min_{i}\{|r_{HT}^{i} - r_{ET}^{i}|\} \tag{3}$$

C_4: The fourth feature is the standard deviation of the difference between ET and HT radius.

C_5: The fifth feature is Mean Relative Tremor (MRT). This feature was proposed by Pereira *et al.* in [14] to measure the amount of tremor of a subject's HT. It is basically the mean difference between the radius of a given sample point and its d left-nearest neighbors. This feature is calculated according the formulation as follows:

$$MRT = \frac{1}{n-d}\sum_{i=d}^{n}\left|r_{ET}^{i} - r_{ET}^{i-d+1}\right| \tag{4}$$

where d is the displacement of the sample points used to compute the radius difference. The following three features are computed based on the relative tremor $\left|r_{ET}^{i} - r_{ET}^{i-d+1}\right|$

C_6: The sixth features denotes the Maximum ET;

C_7: The seventh features denotes the Minimum ET;

C_8: The eighth features denotes the standard derivation of ET values;

C_9: The ninth feature denotes the number of times the difference between HT and ET radius changes from negative to positive, or vice-versa.

III. PROBLEMS AND PROPOSED SOLUTIONS

In this section, we discuss two main problems that are concerned with HandPD dataset. The first problem is the imbalance nature of the data and its impact on the constructed machine learning models. It has been reported in literature that when machine learning models are trained using imbalanced data, the models show biased performance by ignoring the minority class and favoring the majority class [20]. It is due to the fact that the minority class instances occur infrequently during training process, hence the predictions about minority class are also rare, undiscovered or ignored [21]. Consequently, test instances belonging to the minority class are misclassified more often than those belonging to the majority class [21]. In case of binary classification like PD detection, a model will show high rate of sensitivity (if patient class is majority class, like in case of HandPD data) and low rate of specificity (when healthy subjects are minority class). Such performance clearly reflects the biased nature of a model towards majority class. However, in previous studies conducted on HandPD, this biased behavior of machine learning models was ignored and no measures were taken into account to alleviate the bias against performance caused by imbalance data.

In literature, different methods have been used to deal with imbalanced data [22]. The commonly used method is resampling technique. Further, the resampling method includes two methods *i.e.* undersampling and oversampling. In oversampling, the minority class samples are duplicated to balance the size of each class in training data. In undersampling, some majority class samples are removed to balance the size of each class during training process. Hence, when a model is trained on balanced data, it is supposed to show unbiased behavior. In literature, different types of undersampling

methods have been proposed. However, it is reported that random undersampling is the simplest method and has shown similar performance to other complex methods [22]. Thus, in this paper, we utilize random undersampling method to balance the training process and alleviate the baisedness in the constructed models. In each iteration/fold of a cross validation experiment, the random undersampling method randomly removes subjects from the larger class until the training data is balanced. Thus, optimizing or balancing the training process. It is important to note, that the resampling methods are only applied to the training data during each iteration of cross validation and not to the overall data before cross validation.

To alleviate the problem of low rate of PD detection accuracy, we develop a cascaded learning system. Previous studies conducted on the hand written drawings dataset for PD detection discusses only feature extraction and classification methods. However, in data mining and machine learning, feature selection methods are usually exploited to improve performance of predictive models or to reduce their complexities. Motivated by this fact, in this paper we develop a cascaded learning system for PD detection based on hand written drawings data. The proposed system cascades Chi2 model with Adaboost ensemble model in order to improve performance of the Adaboost model and to reduce its complexity. The Chi2 is used for features ranking while the Adaboost model is used as a predictive model to predict presence or absence of PD. To understand the working of the cascaded learning system, we discuss the two basic models *i.e.* Chi2 and Adaboost and their cascade as follows:

Boosting is an ensemble learning method that tries to arrive at a strong learning model by combining the learning capabilities of weak learners. Adaptive boosting or Adaboost model is the first practical boosting ensemble model that was proposed by Freund and Schapire [23]. In other words, Adaboost model converts a set of weak classifiers or estimators into a strong one. It combines the output of other learning algorithms (weak learners or estimators) by evaluating their weighted sum that denotes the final output of the boosting ensemble model. The final equation of Adaboost model for classification can be formulated as follows:

$$G(x) = sign\left(\sum_{m=1}^{M} \beta_m g_m(x)\right) \tag{5}$$

where g_m stands for the m-th weak classifier and β_m is its corresponding weight. It is evident from (5) that Adaboost model is the weighted combination of M weak learners or estimators. Details about working and formulation of Adaboost model can be found in [24], [25]. In this paper, we briefly discuss the formulation of the Adaboost model as follows:

For a given dataset having n instances and binary labels (*i.e.* considering the case of binary classification like the one considered in this paper), the feature vector x and class label y can be denoted as $x_i \in R^d$, $y_i \in \{-1, 1\}$ where -1 denotes negative class (like absence of PD) and $+1$ denotes positive class (like presence of PD). In the first step, weights for each data point are initialized as follows:

$$w(x_i, y_i) = \frac{1}{n}, \quad i = 1, 2, 3, \ldots, n \tag{6}$$

In next step, we iterate from m = 1 to M and fitting weak classifiers to the dataset and select the one that yields lowest weighted classification error.

$$e_m = E_{wm}[1_{y \neq g(x)}] \tag{7}$$

In next step, the weight for the m-th weak classifier or estimator is calculated as follows:

$$\beta_m = \frac{1}{2} \ln\left(\frac{1 - e_m}{e_m}\right) \tag{8}$$

Any classifier (weak estimator) having accuracy higher than 50%, will have positive weight. Additionally, more accurate classifiers will have higher weights. However, classifiers that have less than 50% accuracy will have negative weights. Predictions of such classifiers are combined by Adaboost by flipping their sign. In this way, a classifier with 30% can be turned to yield 70% owing to the sign flipping of its prediction. The only unwanted classifiers are those having exact 50% which contributes nothing to the final prediction.

In next step, the weights of each data point are updated as follows:

$$w_{m+1}(x_i, y_i) = \frac{w_m(x_i, y_i) \exp[-\beta_m y_i g_m(x_i)]}{Z_m} \tag{9}$$

where Z_m is a normalization factor that is used to make sum of all instance weights equal to 1. Furthermore, from (9), it is clear that the "exp" term will always be larger than 1 if a misclassified case is from a positive weighted classifier (*i.e.* β_m is positive and $y * g$ is always negative). That is the misclassified cases will be updated with larger weights after each iteration. The same idea is applied to negative weighted classifiers with the only difference that the original correct classifications would become misclassifications after flipping the sign. Finally, after M iterations, the Adaboost model will obtain final prediction by summing up the weighted prediction of each classifier (*i.e.* weak estimator).

In this paper, we implemented Adaboost ensemble model in scikit-learn library of Python software package [26]. In the rest of the paper, the hyperparameter of the Adaboost model *i.e.* the number of estimators used to construct the final ensemble model will be denoted by N_{est}. Additionally, the base estimator used is decision tree classifier.

In this study, to find out the most relevant features in the feature space, we rank the features through Chi2 statistical test. Chi2 test basically measures the dependency between a feature and a class, thus, successfully finds out features that are more relevant for a given dataset. Hence, we can eliminate those features from the feature space that are irrelevant for classification. The first step in the process of Chi2 test is the construction of TABLE 1.

In the table, ω represents the number of instances (both positive and negative) that accommodate feature F while

TABLE 1. Table for calculating Chi2 score.

	Positive Class	Negative Class	Total
Feature F occurs	α	β	$\alpha + \beta = \omega$
Feature F does not occur	ν	γ	$\nu + \gamma = \sigma - \omega$
Total	$\alpha + \nu = \rho$	$\beta + \gamma = \sigma - \rho$	σ

$\sigma - \omega$ represents the number of instances that do not contain feature F. Similarly, ρ shows the number of positive instances and $\sigma - \rho$ denotes the number of negative instance.

In order to evaluate that how much the expected count *i.e.* X and the observed count *i.e.* D derivate from each other, we use Chi2 test. Let α, β, ν and γ denote the observed values, and X_α, X_β, X_ν and X_α express the expected values then the expected values based on the null hypothesis that the two events are sovereign can be calculated as

$$X_\alpha = (\alpha + \beta)\frac{\alpha + \beta}{\sigma} \tag{10}$$

Similar to (10), X_β, X_ν and X_α can also be calculated. From general formulation of Chi2 test, we have

$$\chi^2 = \frac{1}{d}\sum_{k=1}^{n}\frac{(D_k - X_k)^2}{X_k} \tag{11}$$

$$\chi^2 = \frac{(\alpha - X_i)^2}{X_i} + \frac{(\beta - X_j)^2}{X_j} + \frac{(\nu - X_k)^2}{X_k} + \frac{(\gamma - X_l)^2}{X_l} \tag{12}$$

After simplification, (12) gets the form

$$\chi^2 = \frac{\sigma(\alpha\sigma - \omega\rho)^2}{\rho(\sigma - \rho)(\sigma - \omega)} \tag{13}$$

After features ranking based on the Chi2 test score (denoted by χ^2 in (13)), next we need to select the optimal number of features out of the ranked features. This is explained in the context of the proposed cascaded learning system discussed as follows:

In this paper, we cascade the two models discussed above *i.e.* Chi2 model and Adaboost model. In order to obtain better performance, we need to optimize the two models. The optimization of the Chi2 model means searching the optimal subset of the ranked features that are generated by the Chi2 model. While optimization of Adaboost model means to search optimal number of estimators *i.e.* N_{est} that would yield better classification performance for each subset of features. In order to meet this objective, we sort features in ascending order according to features importance reflected by the Chi2 test score. Thus, the first feature is the feature with highest Chi2 test score *i.e.* the most important feature. The second feature is the feature that contains second highest Chi2 test score *i.e.* it is the second most important or relevant feature and so on.

After features preprocessing, we consider the first subset of features by considering only one feature having highest importance *i.e.* S = 1, where S denotes the size of subset of features. The subset of features is applied to the Adaboost ensemble model. The Adaboost model has its own hyper-parameter N_{est} *i.e.* the number of estimators used. In order to obtain better classification performance, we search the optimal value of N_{est} by using exhaustive search strategy. Thus, the best performance obtained for the first subset of features is obtained under optimized Adaboost model and the performance is noted. In next iteration, another subset of features is constructed by addition of another feature having second highest importance into the previous subset of features *i.e.* we construct subset of features with S = 2. The subset of features is applied to the Adaboost ensemble model and the optimized version of Adaboost ensemble model is obtained by using the same exhaustive search strategy. The performance of this subset of features is also noted. The same process is repeated until all the ranked features are added to the subset of features. Finally, we report the subset of features as optimal subset and the Adaboost model as optimal Adaboost model that yield best performance.

IV. VALIDATION AND EVALUATION

In order to validate the effectiveness of the proposed cascaded system, we utilize stratified k-fold validation scheme with value of k = 4 and k = 5. In the first three experiments, we use k = 5 and in the last experiment k = 4 is utilized. To evaluate the effectiveness of the proposed cascaded system, we tested it against five different evaluation metrics. These metrics include accuracy, sensitivity, specificity, F-score or F-measure and Mathews correlation coefficient (*MCC*). However, conventional accuracy metric fails to reflect the true behavior of a model and this fact is also demonstrated in experiment 1 of the section V of this paper. Thus, we propose to use the balanced accuracy metric which better reflects the true behavior of the constructed models [27]–[29]. In previous studies, Pereira *et al.* [15] used similar accuracy metric (named global accuracy in [15]) that was proposed by Papa *et al.* in [30]. This accuracy metric is also a good choice for reflecting or measuring the true behaviour of a model when it is trained on imbalanced data. In the below formulation, *ACC* denotes the conventionally used accuracy metric and ACC_{bal} denotes the balanced accuracy metric. The formulation of these evaluation metrics is given as follows:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{14}$$

where *TP* denotes the number of true positives, *FP* denotes the number of false positives, *TN* denotes the number of true negatives and *FN* denotes the number of false negatives.

$$Sen = \frac{TP}{TP + FN} \tag{15}$$

$$Spec = \frac{TN}{TN + FP} \tag{16}$$

$$ACC_{bal} = \frac{Sen + Spec}{2} \tag{17}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{18}$$

$$F = \frac{2TP}{2TP + FP + FN} \tag{19}$$

F denotes F-score also known as F-measure, or F_1 score in statistical analysis of binary classification. F returns value between 0 and 1 where 1 indicates perfect predictions and 0 means worst predictions. MCC is used to measure a test's accuracy. MCC can have a value between -1 and 1 where 1 indicates perfect predictions and -1 means worst predictions.

In order to demonstrate the advantages of using balanced accuracy metric, consider a case of 100 subjects with 90 PD patients and 10 healthy subjects. If we construct a model that will always predict a subject to be PD patient, then it will yield 100% of sensitivity but 0% of specificity and conventional accuracy of 90%. However, the balance accuracy will be 50%. Thus, it is evident that the balanced accuracy metric reflects the true behaviour of the constructed model as the model can detect only one class but completely failed to detect the second class. However, the conventional accuracy failed to reflect the true behaviour of the constructed model.

V. EXPERIMENTAL RESULTS AND DISCUSSION

In this section, four different types of experimental settings have been established for simulating the proposed method and validating its effectiveness. The first three experiments use 5-fold stratified cross validation while the fourth experiment uses 4-fold. In the first experiment, the impact of imbalanced HandPD on constructed models is demonstrated. In the second experiment the effectiveness of the proposed cascaded system is validated by comparing it with other similar models while simulating it on spirals data. In the third experiment, we further validate the effectiveness of the proposed method by simulating it on meanders data. All the machine learning models including Adaboost, SVM, GNB, DT, LDA and KNN were implemented using scikit-learn framework of Python package. These models were optimized using grid search algorithm. The main reason for choosing these models is that they have shown state-of-the-art performance on many biomedical and health informatics problems. Authors can find the source code on github using "reliable Parkinson's detection".

A. EXPERIMENT NO 1: HIGHLIGHTING AND DEMONSTRATING IMPACT OF IMBALANCED DATA ON THE CONSTRUCTED MODELS

1) EXPERIMENTS USING SPIRAL DRAWINGS DATA

In this experiment, we demonstrate the impact of imbalanced data on the fitted or constructed models. The results shown in TABLE 2 clearly show that machine learning models are sensitive to imbalanced data. The constructed models are biased towards majority class. For example, it can be seen that when the models are trained using imbalanced data, we obtain high rate of sensitivity and low rate of specificity (highlighted in yellow in the table). It can also be seen that the conventionally used accuracy metric cannot reflect the biasedness in the constructed models as the metric shows good performance (see red text). However, the balanced accuracy truly reflects the original performance of the constructed models. To avoid the problem of biasedness in constructed model, we take measures to balance the training data. As discussed above, we use random undersampling method to balance the size of each class in the training data. After balancing the training process, it is evident from the table that the biasedness in the constructed models have been alleviated *i.e.* the models no more show biased performance as can be seen from the sensitivity and specificity rates highlighted in green in the table. Furthermore, after balancing the training process, the conventional accuracy metric is approaching the balanced accuracy metric (see the red and cyan color text).

TABLE 2. Demonstration of impact of the imbalanced HandPD (Spirals) data on the constructed models. BT: Balanced training, IBT: Imbalance training, ACC_{bal}: Balanced accuracy, ACC: Conventional accuracy metric, MCC: Mathews Correlation Coefficient, Sens[itivity] and Spec[cificity]. F: F-score or F-measure.

Training	Model	ACC	ACC_{bal}	Sen(%)	Spec(%)	F	MCC
IBT	LDA	79.89	53.86	96.62	11.11	0.885	0.142
IBT	KNN	81.25	57.33	96.62	18.05	0.892	0.240
IBT	DT	74.20	59.77	83.44	36.11	0.838	0.192
IBT	GNB	40.46	58.80	28.71	88.88	0.437	0.160
BT	LDA	59.53	58.53	60.13	56.94	0.704	0.136
BT	KNN	57.82	60.67	56.08	65.27	0.681	0.169
BT	DT	59.01	57.15	60.13	54.16	0.702	0.114
BT	GNB	42.19	58.23	31.75	84.72	0.468	0.144

2) EXPERIMENTS USING MEANDER DRAWINGS DATA

In this experiment, we further validate the above findings by performing similar experiments using meander drawings. Again, we demonstrate the impact of imbalanced data on the fitted or constructed models using meander drawings. The results shown in TABLE 3 for meanders data also show that machine learning models are sensitive to imbalanced data. The constructed models are biased towards majority class (evident from sensitivity and specificity rates highlighted in yellow in TABLE 3). Again, we randomly undersample the majority class in the training data to balance the size of each class. After, balancing the training process, it is evident from the table that the biased nature of the model has been alleviated as can be seen from the sensitivity and specificity rates highlighted in green in the table. Additionally, it is important to note from TABLE 2 and TABLE 3 that data balance or imbalance has little impact on GNB model.

B. EXPERIMENT NO 2: DEVELOPMENT OF THE PROPOSED CASCADED LEARNING SYSTEM TO IMPROVE THE PD DETECTION USING SPIRALS DATA AND ITS COMPARISON WITH OTHER SIMILAR MODELS

In this experiment, we develop the proposed cascaded learning system *i.e.* Chi2-Adaboost. In order to validate the effectiveness of the proposed cascaded system, we also develop

TABLE 3. Demonstration of impact of the imbalanced HandPD (Meanders) data on the constructed models. BT: Balanced training, IBT: Imbalance training, ACC_{bal}: Balanced accuracy, ACC: Conventional accuracy metric, MCC: Mathews Correlation Coefficient, Sens[itivity] and Spec[cificity]. F: F-score or F-measure.

Training	Model	ACC	ACC_{bal}	Sen(%)	Spec(%)	F	MCC
IBT	LDA	83.99	64.80	96.28	33.33	0.906	0.400
IBT	KNN	81.80	56.62	97.97	15.27	0.896	0.250
IBT	DT	74.22	60.82	82.77	38.88	0.837	0.209
IBT	GNB	61.77	63.04	60.81	65.27	0.718	0.208
BT	LDA	75.85	74.45	76.68	72.22	0.836	0.413
BT	KNN	64.66	60.69	67.22	54.16	0.753	0.175
BT	DT	61.94	62.16	61.82	62.50	0.723	0.194
BT	GNB	77.71	59.34	89.52	29.16	0.866	0.212

TABLE 4. Development of the proposed cascaded learning system using HandPD (Spirals) data and its comparison with other similar models. S: Size of subset of features, ACC_{bal}: Balanced accuracy, MCC: Mathews Correlation Coefficient, Sens[itivity] and Spec[cificity]. F: F-score or F-measure.

Method	S	ACC_{bal}	Sen(%)	Spec(%)	F	MCC
Chi2-GNB	1	65.46	48.49	81.94	0.638	0.247
Chi2-DT	2	62.03	58.78	65.27	0.703	0.191
Chi2-LDA	1	68.33	54.72	81.94	0.687	0.291
Chi2-KNN	2	64.63	58.44	70.83	0.706	0.232
Chi2-SVM(Lin)	5	51.40	61.14	41.66	0.697	0.022
Chi2-SVM(RBF)	2	64.54	77.70	51.38	0.819	0.257
Chi2-Adaboost	**2**	**72.46**	**69.96**	**75.00**	**0.794**	**0.365**

TABLE 5. Development of the proposed cascaded learning system using HandPD (Meanders) data and its comparison with other similar models. S: Size of subset of features, ACC_{bal}: Balanced accuracy, MCC: Mathews Correlation Coefficient, Sens[itivity] and Spec[cificity]. F: F-score or F-measure.

Method	S	ACC_{bal}	Sen(%)	Spec(%)	F	MCC
Chi2-GNB	5	62.57	82.09	43.05	0.837	0.237
Chi2-DT	5	68.20	64.18	72.22	0.750	0.291
Chi2-LDA	8	75.63	79.05	72.22	0.850	0.439
Chi2-KNN	3	62.27	66.21	58.33	0.750	0.200
Chi2-SVM(Lin)	4	57.90	56.08	59.72	0.676	0.125
Chi2-SVM(RBF)	4	59.75	34.79	84.72	0.502	0.167
Chi2-Adaboost	**5**	**78.04**	**68.58**	**87.50**	**0.799**	**0.450**

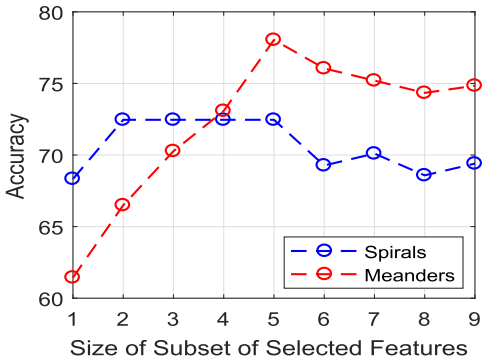


FIGURE 4. Performance of the proposed Chi2-Adaboost model at different subsets of features.

other similar cascaded systems *e.g.*, Chi2 model cascaded with Linear Discriminant Analysis model (LDA), with Gaussian Naive Bayes (GNB), with Decision Tree (DT), with K Nearest Neighbors (KNN), with SVM Linear and with SVM RBF. The results for each of the developed cascaded model are shown in TABLE 4. These results are obtained by simulating the cascaded systems using spiral drawings data. It can be seen that the highest accuracy of 72.46% is achieved by the proposed Chi2-Adaboost model. It is important to note that the proposed method achieved this results using only a small subset of features with S = 2. Hence, the propose method also reduces the complexity of the Adaboost predictive model as training on less number of features will result in lower training time. It is important to note that for the optimal subset of features with S = 2, the optimized Adaboost model was obtained at $N_{est} = 2$ using grid search algorithm.

C. EXPERIMENT NO 3: DEVELOPMENT OF THE PROPOSED CASCADED LEARNING SYSTEM TO IMPROVE THE PD DETECTION USING MEANDERS DATA AND ITS COMPARISON WITH OTHER SIMILAR MODELS

In this experiment, we develop the proposed cascaded learning system for meander drawings data. In order to validate the effectiveness of the proposed cascaded system, we also develop other similar cascaded systems *i.e.*, Chi2-GNB, Chi2-DT, Chi2-LDA, Chi2-KNN, Chi2-SVM(Lin) and Chi2-SVM(RBF). The results for each of the developed cascaded model are shown in TABLE 5. It can be seen that the highest accuracy of 78.04% is achieved by the proposed Chi2-Adaboost model. It is important to note that the proposed method achieved this results using only a small subset of features with S = 5. Hence, the proposed method also reduces the complexity of the Adaboost predictive model as training on less number of features will result in lower training time. Moreover, for the optimal subset of features with S = 5, the optimized Adaboost model was obtained at $N_{est} = 37$ using grid search algorithm. From the experimental results, it is clear that the optimal subset of features contains complimentary information about PD compared to the full features space.

The performance of the proposed method at different subsets of features (for both spirals and meanders data) is given in FIGURE 4. It can be seen that for spiral data, the best performance of 72.46% is obtained using subset of features with sizes 2, 3, 4 and 5. However, we reported $S = 2$ *i.e.* subset of features with size 2 as optimal as it would construct a less complex model (in terms of time complexity). Moreover, the selected features for each subset are reported in TABLE 6 and 7. In the tables, a feature having value of True means the feature is selected while a feature with value False means it is rejected.

D. COMPARATIVE STUDY WITH CONVENTIONAL ADABOOST MODEL

In this subsection, to further validate the effectiveness of the proposed cascaded system, we compare performance of the proposed cascaded system with conventional Adaboost model. For spiral data, we simulated conventional Adaboost

TABLE 6. Details about selected features for spiral data.

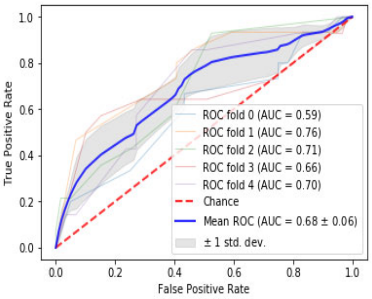
Size	Features								
S	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9
S=1	True	False	False	False	False	False	False	False	False
S=2	True	False	True	False	False	False	False	False	False
S=3	True	True	True	False	False	False	False	False	False
S=4	True	True	True	True	False	False	False	False	False
S=5	True	True	True	True	False	False	False	True	False
S=6	True	True	True	True	False	True	False	True	False
S=7	True	True	True	True	True	True	False	True	False
S=8	True	True	True	True	True	True	False	True	True

TABLE 7. Details about selected features for meander data.

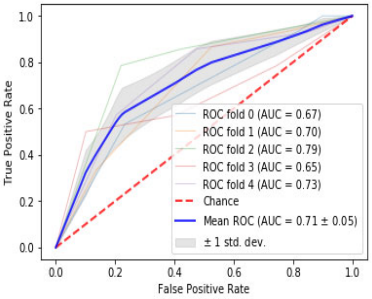
Size	Features								
S	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9
S=1	False	False	True	False	False	False	False	False	False
S=2	True	False	True	False	False	False	False	False	False
S=3	True	False	True	False	False	False	False	True	False
S=4	True	True	True	False	False	False	False	True	False
S=5	True	True	True	True	False	False	False	True	False
S=6	True	True	True	True	False	True	False	True	False
S=7	True	True	True	True	True	True	False	True	False
S=8	True	True	True	True	True	True	True	True	False

model and it achieved 69.40% accuracy on optimal $N_{est} = 4$ while the proposed cascaded system achieved 72.46% accuracy. Hence, it is proved that the proposed cascaded method has improved the strength of conventional Adaboost model by 3.06% for spiral data. Moreover, we also simulated the conventional Adaboost model on meander data and achieved 74.80% accuracy on optimal $N_{est} = 55$ while the proposed cascaded system obtained 78.04% accuracy. Hence, it is evidently clear that the proposed method improved the strength of conventional Adaboost model by 3.24% for meander data.

To further validate the fact that the proposed method improves the strength of conventional Adaboost model, we utilized two more evaluation metrics *i.e.* ROC chart and area under the curve (AUC). The ROC chart for conventional Adaboost model is shown in FIGURE 5 (a) and the ROC chart for the proposed cascaded system is shown in FIGURE 5 (b). These ROC charts are obtained considering spiral data. Similarly for meander data, the ROC chart for conventional Adaboost model is shown in FIGURE 6 (a) and the ROC chart for the proposed system is shown in FIGURE 6 (b). Thus, it is also evident from the ROC charts that the proposed method has improved the strength of conventional Adaboost model as the ROC charts for the cascaded system have more AUC compared to ROC charts of conventional Adaboost model. It is important to note that the improvement in performance of conventional Adaboost model is due to the fact that Chi2 model successfully eliminate some irrelevant features from the feature space before their application to the Adaboost model. If we train conventional Adaboost model without features refinement through Chi2 model, it will learn some noisy patterns from the irrelevant features during training process, thus, will show degradation in performance on the testing data.

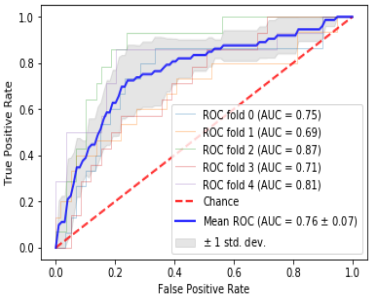


(a) ROC chart of conventional Adaboost model

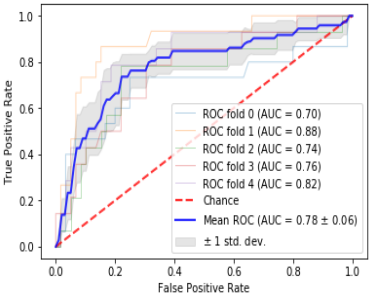


(b) ROC chart of the proposed Chi2-Adaboost cascaded system

FIGURE 5. ROC charts using conventional Adaboost model and the proposed cascaded Chi2-Adaboost for spiral data.



(a) ROC chart of conventional Adaboost model



(b) ROC chart of the proposed Chi2-Adaboost cascaded system

FIGURE 6. ROC charts using conventional Adaboost model and the proposed cascaded Chi2-Adaboost for meander data.

TABLE 8. Performance of the proposed cascaded system on spiral data using 4-fold cross validation.

Method	S	ACC_{bal}	Sen(%)	Spec(%)	F	MCC
Chi2-GNB	3	63.25	45.49	80.55	0.609	0.213
Chi2-DT	3	61.31	60.13	62.5	0.710	0.180
Chi2-LDA	7	68.20	64.18	72.22	0.750	0.291
Chi2-KNN	2	66.06	55.74	76.38	0.690	0.254
Chi2-SVM(Lin)	3	51.91	49.66	54.16	0.617	0.030
Chi2-SVM(RBF)	1	62.93	56.41	69.44	0.688	0.205
Chi2-Adaboost	**2**	**69.51**	**69.59**	**69.44**	**0.786**	**0.318**

TABLE 9. Performance of the proposed cascaded system on meander data using 4-fold cross validation.

Method	S	ACC_{bal}	Sen(%)	Spec(%)	F	MCC
Chi2-GNB	5	68.97	71.28	66.66	0.794	0.313
Chi2-DT	4	65.09	63.51	66.66	0.740	0.242
Chi2-LDA	7	74.00	84.12	63.88	0.872	0.438
Chi2-KNN	3	63.85	65.20	62.50	0.748	0.224
Chi2-SVM(Lin)	4	58.65	27.02	90.27	0.417	0.161
Chi2-SVM(RBF)	1	54.52	11.82	97.22	0.210	0.119
Chi2-Adaboost	**8**	**76.44**	**70.94**	**81.94**	**0.809**	**0.429**

E. EXPERIMENT NO 4: PERFORMANCE OF THE PROPOSED CASCADED METHOD USING 4-FOLD CROSS VALIDATION

In this subsection, we check the performance of the proposed cascaded system by using the value of k = 4 for the stratified cross validation and utilizing undersampling with replacement method for balancing the training process. The main objective of this experiment is to further validate the effectiveness of the proposed method by using undersampling with replacement method and to reduce subject dependence in training and testing datasets. The simulation results for spiral data are reported in TABLE 8 and the results for meander data are given in TABLE 9. Moreover, it was also observed that for the spirals data, conventional Adaboost yielded 67.58% accuracy (the optimal value of $N_{est} = 2$ is obtained for spiral data). For meanders data, the conventional Adaboost model obtained 75.05% accuracy (the optimal value of $N_{est} = 19$ is obtained for the meanders data). Thus, it is further validated that the proposed cascaded model improves the strength of conventional Adaboost model.

VI. CONCLUSION AND FUTURE WORK

In this paper, we considered the problem of PD detection based on handwritten data. The data under consideration was highly imbalanced in nature. We experimentally demonstrated the biasedness in the machine learning models that is caused by the imbalanced data. It was shown that when machine learning models are trained on imbalanced data, their performance is biased towards the majority class in the data. Hence, for the PD detection problem, we observed high rate of sensitivity and low rate of specificity as the patient class was in majority and healthy class was in minority. To alleviate the biasedness in the constructed models, we utilized random undersampling method. After the optimization or balancing of the training process through random undersampling method, unbiased models were developed. Moreover, to improve PD detection accuracy, feature selection method was integrated with machine learning methods. Thus, a cascaded learning system namely Chi2-Adaboost was developed. It was shown that the proposed system outperformed six similar cascaded learning systems. Additionally, it was also observed that the proposed cascaded system improved the performance of a conventional Adaboost model by 3.3%.

Although, in this study, the problem of biasedness in constructed models was avoided and an unbiased cascaded model was developed that improved the PD detection accuracy as well as reduced the complexity of machine learning models by reducing the number of features. However, the obtained accuracy still needs considerable amount of improvement. This is a limitation of this study. In future studies, we need to develop more robust models that can improve the PD detection accuracy while maintaining the unbiased behaviour of the constructed models. This can be achieved by integrating feature selection methods with deep learning models.

REFERENCES

[1] A. Benba, A. Jilbab, and A. Hammouch, "Analysis of multiple types of voice recordings in cepstral domain using MFCC for discriminating between patients with Parkinson's disease and healthy people," *Int. J. Speech Technol.*, vol. 19, no. 3, pp. 449–456, 2016.

[2] S. K. van den Eeden, C. M. Tanner, A. L. Bernstein, R. D. Fross, A. Leimpeter, D. A. Bloch, and L. M. Nelson, "Incidence of parkinson's disease: Variation by age, gender, and race/ethnicity," *Amer. J. Epidemiol.*, vol. 157, no. 11, pp. 1015–1022, 2003.

[3] L. Cunningham, S. Mason, C. Nugent, G. Moore, D. Finlay, and D. Craig, "Home-based monitoring and assessment of Parkinson's disease," *IEEE Trans. Inf. Technol. Biomed.*, vol. 15, no. 1, pp. 47–53, Nov. 2011.

[4] Z. Dastgheib, B. Lithgow, and Z. Moussavi, "Diagnosis of Parkinson's disease using electrovestibulography," *Med. Biol. Eng. Comput.*, vol. 50, no. 5, pp. 483–491, 2012.

[5] G. Rigas, A. T. Tzallas, M. G. Tsipouras, P. Bougia, E. E. Tripoliti, D. Baga, D. I. Fotiadis, S. G. Tsouli, and S. Konitsiotis, "Assessment of tremor activity in the Parkinson's disease using a set of wearable sensors," *IEEE Trans. Inf. Technol. Biomed.*, vol. 16, no. 3, pp. 478–487, May 2012.

[6] M. A. Little, P. E. McSharry, E. J. Hunter, J. Spielman, and L. O. Ramig, "Suitability of dysphonia measurements for telemonitoring of Parkinson's disease," *IEEE Trans. Biomed. Eng.*, vol. 56, no. 4, pp. 1015–1022, Apr. 2009.

[7] L. Ali, C. Zhu, M. Zhou, and Y. Liu, "Early diagnosis of parkinson's disease from multiple voice recordings by simultaneous sample and feature selection," *Expert Syst. Appl.*, vol. 137, pp. 22–28, Dec. 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S095741741930452X

[8] R. Das, "A comparison of multiple classification methods for diagnosis of Parkinson disease," *Expert Syst. Appl.*, vol. 37, no. 2, pp. 1568–1572, 2010.

[9] L. Parisi, N. RaviChandran, and M. L. Manaog, "Feature-driven machine learning to improve early diagnosis of Parkinson's disease," *Expert Syst. Appl.*, vol. 110, pp. 182–190, Nov. 2018.

[10] L. Naranjo, C. J. Pérez, J. Martín, and Y. Campos-Roca, "A two-stage variable selection and classification approach for Parkinson's disease detection by using voice recording replications," *Comput. Methods Programs Biomed.*, vol. 142, pp. 147–156, Apr. 2017.

[11] P. Drotár, J. Mekyska, I. Rektorová, L. Masarová, Z. Smékal, and M. Faundez-Zanuy, "Analysis of in-air movement in handwriting: A novel marker for parkinson's disease," *Comput. Methods Programs Biomed.*, vol. 117, no. 3, pp. 405–411, 2014.

[12] S. Rosenblum, M. Samuel, S. Zlotnik, I. Erikh, and I. Schlesinger, "Handwriting as an objective tool for Parkinson's disease diagnosis," *J. Neurol.*, vol. 260, no. 9, pp. 2357–2361, 2013.

[13] P. Drotár, J. Mekyska, I. Rektorová, L. Masarová, Z. Smékal, and M. Faundez-Zanuy, "Evaluation of handwriting kinematics and pressure for differential diagnosis of Parkinson's disease," *Artif. Intell. Med.*, vol. 67, pp. 39–46, Feb. 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0933365716000063

[14] C. R. Pereira, D. R. Pereira, F. A. da Silva, C. Hook, S. A. Weber, L. A. Pereira, and J. P. Papa, "A step towards the automated diagnosis of Parkinson's disease: Analyzing handwriting movements," in *Proc. IEEE 28th Int. Symp. Comput.-Based Med. Syst.*, Jun. 2015, pp. 171–176.

[15] C. R. Pereira, D. R. Pereira, F. A. Silva, J. P. Masieiro, S. A. Weber, C. Hook, and J. P. Papa, "A new computer vision-based approach to aid the diagnosis of Parkinson's disease," *Comput. Methods Programs Biomed.*, vol. 136, pp. 79–88, Nov. 2016.

[16] C. R. Pereira, L. A. Passos, R. R. Lopes, S. A. Weber, C. Hook, and J. P. Papa, "Parkinson's disease identification using restricted boltzmann machines," in *Proc. Int. Conf. Comput. Anal. Images Patterns*. Springer, 2017, pp. 70–80.

[17] C. R. Pereira, S. A. T. Weber, C. Hook, G. H. Rosa, and J. P. Papa, "Deep learning-aided parkinson's disease diagnosis from handwritten dynamics," in *Proc. 29th SIBGRAPI Conf. Graph., Patterns Images (SIBGRAPI)*, 2016, pp. 340–346.

[18] C. R. Pereira, D. R. Pereira, S. A. Weber, C. Hook, V. H. C. de Albuquerque, and J. P. Papa, "A survey on computer-assisted Parkinson's disease diagnosis," *Artif. Intell. Med.*, vol. 95, pp. 48–63, Apr. 2019.

[19] C. R. Pereira, D. R. Pereira, F. A. Silva, J. P. Masieiro, S. A. Weber, C. Hook, and J. P. Papa. (2016). *Handpd Dataset*. Accessed: Jan. 15, 2019. [Online]. Available: http://wwwp.fc.unesp.br/~papa/pub/datasets/Handpd/

[20] Y. Wang, A.-N. Wang, Q. Ai, and H.-J. Sun, "An adaptive kernel-based weighted extreme learning machine approach for effective detection of Parkinson's disease," *Biomed. Signal Process. Control*, vol. 38, pp. 400–410, Sep. 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1746809417301271

[21] Y. Sun, A. K. Wong, and M. S. Kamel, "Classification of imbalanced data: A review," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 23, no. 4, pp. 687–719, 2009. doi: 10.1142/S0218001409007326.

[22] N. Japkowicz, "The class imbalance problem: Significance and strategies," in *Proc. Int. Conf. Artif. Intell.*, 2000, pp. 1–7.

[23] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 119–139, Aug. 1997.

[24] X. Tang, Z. Ou, T. Su, and P. Zhao, "Cascade AdaBoost classifiers with stage features optimization for cellular phone embedded face detection system," in *Proc. Int. Conf. Natural Comput.* Springer, 2005, pp. 688–697.

[25] S. Prabhakar and H. Rajaguru, "AdaBoost classifier with dimensionality reduction techniques for epilepsy classification from EEG," in *Precision Medicine Powered by pHealth and Connected Health*. Springer, 2018, pp. 185–189.

[26] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.

[27] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The balanced accuracy and its posterior distribution," in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 3121–3124.

[28] D. R. Velez, B. C. White, A. A. Motsinger, W. S. Bush, M. D. Ritchie, S. M. Williams, and J. H. Moore, "A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction," *Genet. Epidemiol.*, vol. 31, no. 4, pp. 306–315, 2010.

[29] T. Tong, C. Ledig, R. Guerrero, A. Schuh, J. Koikkalainen, A. Tolonen, H. Rhodius, F. Barkhof, B. Tijms, and A. W. Lemstra, "Five-class differential diagnostics of neurodegenerative diseases using random undersampling boosting," *NeuroImage, Clin.*, vol. 15, pp. 613–624, Jan. 2017.

[30] J. P. Papa, A. X. Falcão, and C. T. N. Suzuki, "Supervised pattern classification based on optimum-path forest," *Int. J. Imag. Syst. Technol.*, vol. 19, no. 2, pp. 120–131, 2009.