Determining the Efficiency of LSVT Treatment in Patients with Parkinson's Disease Using Different Feature Selection & Classification Algorithms

Merna Atef, Youssef Gamal, Eman Moustafa, and Peter Emad

***Abstract*—Speech disorders such as dysphonia and dysarthria represent an early and common manifestation of Parkinson's disease. Class prediction is an essential task in automatic speech treatment, particularly in Parkinson's disease cases. Many classification experiments have been performed which focus on the automatic detection of Parkinson's disease patients from healthy speakers but results are still not very optimistic. Our aim in this study is to build a Lee Silverman Voice Treatment (LSVT) system for assessing the vocal performance of the patients as acceptable or unacceptable, using Naive Bayes classifier and Sequential Forward Selection (SFS), we were able to reduce the features from f=310, to f=6 and obtained an accuracy of 94.62%, a specificity of 97.26%, and a sensitivity of 94.25%.**

***Index Terms*—classification, Naive Bayes, features reduction, Parkinson's disease, sequential forward selection, machine learning, LSVT.**

I. INTRODUCTION

Parkinson's disease is a neurodegenerative disorder that affects the nerve cells in the brain that produce dopamine, which leads to progressive deterioration of motor function due to loss of these dopamine-producing brain cells [1]. The number of individuals with PD over age 50 in the world's high population countries was between 4.1 and 4.6 million in 2005. This number is expected to double to between 8.7 and 9.3 million by 2030 [2]. The cause of Parkinson's disease is unknown, many people with Parkinson's disease (PD) will experience problems with their voices. About 45% to 89% of patients report speech problems, and more than 30% find these speech problems to be the most debilitating part of the disease. The patient's voice may get softer, breathy, or hoarse, causing others difficulty in hearing what is said, the speech may also be slurred, mumbled or expressed rapidly. The tone of the voice may become monotone, lacking the normal ups and downs. These problems with communication can result in social isolation. Treatment is customized to the individual's needs, as well as the difficulties they are experiencing [3][4]. Speech therapy has been shown to help people with PD. A speech therapist, or a speech language pathologist can evaluate and treat several speech difficulties.

One program with exercises specifically for people with PD is called Lee Silver Voice Treatment (LSVT). LSVT has been shown to be the best behavioral therapy program for both short-term and long-term speech treatment for patients with PD. LSVT program improves phonatory effort and vocal characteristics (loudness, pitch variability, vocal quality), as well as improves speech articulation. The aim of our study is to investigate the potential of using an objective statistical machine learning framework to automatically evaluate sustained vowel phonations as acceptable (a clinician would allow persisting in the speech treatment) or unacceptable (a clinician would not allow persisting in the speech treatment). The ultimate goal is to improve the effectiveness of rehabilitative speech treatment by developing an appropriate algorithm for the LSVT companion system. Using an algorithm to detect how different features contribute to the assessment of the efficiency of the treatment. In the original study, two classification algorithms - Random Forests (RF), and Support Vector Machines (SVM)- were used and the results reached had an accuracy of 90%. So far little studies have been published to assess the efficiency of using a Naïve Bayes classifier algorithm to solve this problem. In this paper, we compared two classifier algorithms (Naive Bayes and K Nearest Neighbors) with various sets of features deduced from numerous feature selection algorithms.

II. MATERIALS & METHODOLOGY

A. Dataset

We used the LSVT dataset from the UCI machine learning repository [5]. The dataset is composed of 126 sustained vowel /a/ phonations' features with 310 dysphonia measures. The signals were measured from 14 patients who have been diagnosed with Parkinson's disease and were undergoing the LSVT rehabilitation program. The data set was originally collected to determine the most parsimonious feature subset which helps to predict the binary response acceptable or unacceptable. The dataset uses the dysphonia measures defined in detail in Tsanas et al. [6], [7], and summarized more recently in Tsanas [8]. Many dysphonia measures rely on the computation of the fundamental frequency (F0). The 310 dysphonia measures used can be divide into groups.

The first group is related to the observation that the vocal fold vibration pattern is nearly periodic in healthy voices, and aperiodic or departs from periodicity in pathological voices. Two of the dysphonia measures used from this category are jitter and shimmer; jitter quantifies F0 deviations, whereas shimmer quantifies deviations in amplitude. Many jitter and

shimmer variants, were investigated and the following measures were used: recurrence period density entropy (RPDE), the pitch period entropy (PPE), the glottal quotient (GQ), and other F0-related measures. The F0-related dysphonia measures include statistical summaries of the F0 distribution, and the difference in the measured F0 of age- and gender-matched healthy controls.

The second general group of dysphonia measures is the signal-to-noise ratio (SNR) type algorithms. They describe the acoustic noise from the incomplete closure of the vocal folds. Examples of this group are: harmonic-to-noise ratio (HNR), detrended fluctuation analysis (DFA), glottal to noise excitation (GNE), vocal fold excitation ratio (VFER), and empirical mode decomposition excitation ratio (EMD-ER) measures. GNE and VFER analyze frequency ranges of sustained vowel phonation in bands of 500 Hz. The frequencies below 2.5 kHz can be treated as signal, and everything above 2.5 kHz can be treated as noise. EMDER forms ratios of signal and noise energies. The empirical mode decomposition (EMD) algorithm decomposes a signal into elementary, linearly superposed signal components with amplitude and frequency contributions. Then, the top (high frequency) components are taken to constitute noise, whereas the lower frequency components are taken to constitute the signal.

They analyzed the F0 time series (also known as F0 contour) using wavelet decomposition with 10 levels of decomposition. Also, the log-transformed F0 time series was used to bring out additional characteristics in dysphonia measures.

Lastly, Mel frequency cepstral coefficients (MFCCs) target the placement of the articulators (collectively referring to the mouth, teeth, tongue, and lips), which is known to be affected in PD. The articulatory features used in the study characterize fluctuations and instability in postural stability of the articulators during sustained vowel phonation [9].

B. Dimensionality Reduction

Dimensionality reduction is the process of reducing the number of random variables (features) under consideration [10]. Dimension reduction techniques reduce noise in the data which minimizes redundancy and maximizes relevance. These techniques increase the classification accuracy as the uninformative and redundant features are removed in the feature selection phase. It can be classified into two groups: feature selection and feature extraction or transformation of features.

The feature selection techniques select a small subset of features from large dataset whereas feature extraction techniques transform features into a lower dimensional feature space [11]. In contrast to feature extraction techniques, feature selection techniques do not alter the original representation of the variables. Thus, they preserve the original semantics of the variables, hence, offering the advantage of interpretability by a domain expert. Feature selection is preferable to feature transformation when the original units and meaning of features are important and the modeling goal is to identify an influential subset. The objectives of feature selection are manifold, the most important ones being: (a) to avoid overfitting and improve model performance, i.e. prediction performance in supervised classification, (b) to provide faster and more cost-effective models and (c) to gain a deeper insight into the underlying processes that generated the data. Feature selection techniques can be organized into three categories, depending on how they combine the feature selection search with the construction of the classification model: filter methods, wrapper methods and embedded methods.

In this paper we used three selection techniques: Correlation-based Selection, Mutual Information (multivariate filter methods), and Forward Subset Selection (wrapper method) [12].

1) Feature Selection

a) Filter Techniques

Filter techniques assess the relevance of features by looking only at the intrinsic properties of the data, they treat the problem of finding a good feature subset independently of the model selection step. They easily scale to very high-dimensional datasets, they are computationally simple and fast, and they are independent of the classification algorithm. As a result, feature selection needs to be performed only once, and then different classifiers can be evaluated. Multivariate filter techniques also incorporate the feature dependencies to some degree [13].

(1) Correlation-based Selection

A feature of a subset is good if it is highly correlated with the class but not much correlated with other features of the class. We used Pearson Correlation Coefficient - developed by Karl Pearson- to find highly correlated features to the class label [14]. The Pearson's correlation coefficient, typically denoted by r, is a measure of the correlation (linear dependence) between two random variables X and Y.

It takes the values in the interval r ∈ [-1,1]. A value of $r = 1$ means that the two variables are in complete agreement. A value of r = -1 means that the two variables take opposite values. Pearson's correlation coefficient between two variables is defined as the covariance of the two variables normalized by the product of their standard deviations.

$$r(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \mu_x)(Y_i - \mu_y)}{\sqrt{\sum_{i=1}^{n}(X_i - \mu_x)^2}\sqrt{\sum_{i=1}^{n}(Y_i - \mu_y)^2}}$$

Where,

$\mu_x = \frac{1}{n}\sum_{i=1}^{n} X_i$, and similarly

$\mu_y = \frac{1}{n}\sum_{i=1}^{n} Y_i$.

This equation gives a value between -1 and +1, where +1 is a maximum positive correlation, 0 is no correlation, and +1 is the strongest negative correlation.

We iterated over the 310 features and selected the 5 features having the maximum absolute correlation coefficient with the response Y.

(2) Mutual Information

Mutual Information (MI) is feature selection algorithm based on ideas from information theory and probabilistic reasoning. It measures how much one random variables tells us about another. It measures the mutual dependence between the two variables. Zero MI means the two variables are independent.

For two discrete variables X and Y whose joint probability distribution is p(X,Y) , the mutual information between them, denoted I(X;Y), is given by

$$I(X;Y) = \sum_{y}\sum_{x} p(X,Y)log\left(\frac{p(X,Y)}{p(X)p(Y)}\right)$$

Where, p(X,Y) is the joint probability density function of X and Y,

$$p(X,Y) = \frac{1}{2\pi\sqrt{1-p^2}}exp\left(-\frac{1}{2(1-p^2)}\left[\frac{(\mathrm{X_i}-\mu_x)^2}{\sigma_x^2} + \frac{(\mathrm{Y_i}-\mu_y)^2}{\sigma_y^2} - 2p\left(\frac{\mathrm{X_i}-\mu_x}{\sigma_x}\right)\left(\frac{\mathrm{Y_i}-\mu_y}{\sigma_y}\right)\right]\right)$$

where p is the correlation between X and Y. and p(X) and p(Y) are the marginal probability density functions of X and Y respectively.

We iterated over the 310 features and selected the 5 features having the maximum MI with the response Y.

b) Wrapper Techniques

Wrapper techniques embed the model hypothesis search within the feature subset search. In this setup, a search procedure in the space of possible feature subsets is defined, and various subsets of features are generated and evaluated. The evaluation of a specific subset of features is obtained by training and testing a specific classification model, rendering this approach tailored to a specific classification algorithm.

To search the space of all feature subsets, a search algorithm is then wrapped around the classification model. Advantages of wrapper approaches include the interaction between feature subset search and model selection, and the ability to consider feature dependencies [15].

(1) Sequential Forward Selection

This method selects a subset of features from the data matrix X that best predict the data in Y by sequentially selecting features until there is no improvement in prediction [16]. This method has two components:

- An objective function, called the criterion, which the method seeks to minimize over all feasible feature subsets. Common criteria are mean squared error (for regression models) and misclassification rate (for classification models). In this paper we used the inaccuracy of prediction as the criterion.
- A sequential search algorithm, which adds (Sequential Forward Selection SFS) or removes (Sequential Backward Selection SBS) features from a candidate subset while evaluating the criterion. Since an exhaustive comparison of the criterion value at all 2n subsets of an n-feature data set is typically infeasible (depending on the size of n and the cost of objective calls), sequential searches move in only one direction, always growing or always shrinking the candidate set [12].

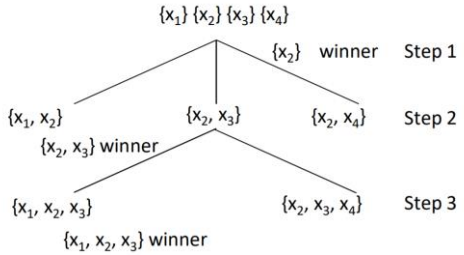

Figure 1 SFS

2) Feature Transformation

We calculated 4 extra features (Energy, 4th power, Nonlinear Energy, Curve Length) from the existing 310 features according to the following equations:

$$Energy = \sum \mathrm{X_i^2}$$

$$4^{\mathrm{th}}Power = \sum \mathrm{X_i^4}$$

$$Nonlinear\ Energy = \sum \text{-}\mathrm{X_i X_{i-2}} + \mathrm{X_{i-1}^2}$$

$$CurveLength = \sum \mathrm{X_i} - \mathrm{X_{i-1}}$$

C. Classifiers

1) Naïve Bayes

This Classification is named after Thomas Bayes (1702-1761), who proposed the Bayes Theorem. The Bayesian Classification represents a supervised learning method as well as a statistical method for classification. It assumes an underlying probabilistic model and allows us to capture uncertainty about the model in a principled way by determining probabilities of the outcomes. It can solve diagnostic and predictive problems.

Bayesian classification provides practical learning algorithms and prior knowledge and observed data can be combined. It also provides a useful perspective for understanding and evaluating many learning algorithms. It calculates explicit probabilities for hypothesis and it is robust to noise in input data [17], [18]. The Bayes rule, says: if you can have a hypothesis Y and data X which bears on the hypothesis, then

$$\mathrm{p(Y|X_i)} = \frac{\mathrm{p(X_i|Y)p(Y)}}{\mathrm{p(X_i)}}$$

Where, $\mathrm{p(X_i)}$ is the independent probability; p(Y) is the prior probability; $\mathrm{p(X_i|Y)}$ is the conditional probability of $\mathrm{X_i}$ given Y (likelihood); $\mathrm{p(Y|X_i)}$ is the conditional probability of Y given $\mathrm{X_i}$ (posterior probability).

In our case $\mathrm{X_i}$ represents the features which we assumed are independent of each other, where $\mathrm{X_i} = \{\mathrm{X_1}, \mathrm{X_2}, \mathrm{X_3}, \dots, \mathrm{X_{310}}\}$ and Y represents the classes, where $\mathrm{Y_i} = \{\mathrm{Y_1}, \mathrm{Y_2}\}$.

We used this classifier as it is simple, easy to implement

and fast. It need less training data and is highly scalable; As it scales linearly with the number of predictors and data points and is very suitable for binary classification problems.

Since our data is continuous, we assumed normal distribution for the features and used the Gaussian distribution formula to calculate the likelihood probability

$$p(X) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(X-\mu_x)^2/2\sigma_x^2}$$

2) K-Nearest Neighbors

K-nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). KNN has been used in statistical estimation and pattern recognition already in the beginning of 1970s as a non-parametric technique.

A case is classified by a majority vote of its neighbors, with the case being assigned to the class most common amongst its KNN measured by a distance function. If K = 1, then the case is simply assigned to the class of its nearest neighbor. We choose the K by trial and error and assigned it a value of 7.

The distance function used is

$$D = \sqrt{\sum_{i=1}^{n} (X_i - Y_i)^2}$$

III. Results and Discussion

We used different feature selection algorithms so that we could reach the optimum outcomes. We also used k-fold cross validation (CV) with K=10; to divide the dataset into different test and training sets to optimize the performance and calculate the mean performance parameters (Accuracy, Sensitivity, and Specificity).

At the beginning we chose K-Nearest Neighbors (KNN) classifier. We tried this classifier with different features (Table 1). First, we used all 310 features, this yielded an accuracy of 68.5%. This was far from our expectations, so we tried to use the four calculated features, and this yielded an accuracy of 66.67%. We tried the given features with the calculated ones, and this gave us accuracy of 68.14%. Still less than the first trial.

Finally, we decided to use feature selection algorithms to reduce the dimensionality of the data. With the Mutual Information selection method (5 features), we achieved an accuracy of 80.26%. With the Correlation-based selection (5 features) we got an accuracy of 81.79%. The best accuracy reached using KNN was 90.51% using Sequential Forward Selection algorithm (SFS)- 4 features.

We aimed to reach even a higher accuracy, so we tried a totally different classifier, Naive Bayes classifier. We followed the same experimental steps as above. We got an accuracy of 62.56% using the 310 features for training. An accuracy of 74.68% and 45.13% were reached using 314 features (original features with the four calculated ones) and the 4 calculated ones respectively.

Using feature selection algorithms (Mutual information and Correlation-based selections – both 5 features), accuracies of 81.92% and 89.04% were reached respectively.

Finally, the best overall accuracy reached was 94.62% using Naïve Bayes classifier and Sequential Forward Selection algorithms (6 features). We also achieved a sensitivity of 94.25% and a specificity of 97.26% in the same test.

C	P	SFS	Corr	MI	310f	314f	4f
NB	M.Acc	94.62	89.0	81.9	62.6	74.7	45.1
	M.Spe	92.26	93.9	95.2	91.1	72.6	76.7
	M.Sen	94.25	84.7	69.6	49.2	-	37.3
KNN	M.Acc	90.51	81.8	80.3	68.5	68.1	66.7
	M.Spe	94	87.4	84.3	74.9	72.9	72.6
	M.Sen	90.5	74.1	79	60.2	59.1	63

Table 1

C: Classification method; P: Evaluation Parameter; Corr: Correlation; M.Acc: Maximum Accuracy; M.Spec: Maximum Specificity; M.Sen: Maximum Sensitivity.

IV. Conclusion

In summary, in this paper we focused on getting the best results from the data we have, to ascertain building the best Lee Silverman Voice Treatment (LSVT) system for assessing the vocal performance; whether through using different classifiers or different algorithms.

The main advantage of our approach is that the number of features can be effectively reduced from p = 310 to p = 6 by SFS method, also reaching the highest accuracy using Naive Bayes classifier and the mentioned method.

The results of extensive testing performed on the LSVT Voice Rehabilitation dataset are: an accuracy of 94.62% with Naive Bayes classification with 6 features only. Those results reveal the advantages of our proposed approach.

V. References

[1] https://www.webmd.com/parkinsons-disease/default.htm

[2] Dorsey, E.R., et al.: Projected number of people with Parkinson disease in the most populous nations, 2005 through 2030. Neurology 68, 384386 (2007) Google Scholar

[3] https://www.medicinenet.com/parkinsonsdisease=article

[4] Lichman, M.: UCI Machine Learning Repository (2013).

[5] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinsons disease symptom severity, J. R. Soc.Interface, vol. 8, pp. 842855, 2011.

[6] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, New nonlinear markers and insights into speech signal degradation for effective tracking of Parkinsons disease symptom severity, in Proc. Int. Symp. Nonlinear TheoryAppl. (NOLTA), Krakow, Poland, Sep. 58, 2010, pp. 457460.

[7] A. Tsanas, Accurate telemonitoring of Parkinsons disease symptom severity using nonlinear speech signal processing and statistical machine learning, D.Phil, Univ. Oxford, Oxford, U.K., 2012.

[8] Tsanas, A., Little, M.A., Fox, C., Ramig, L.O.: Objective automatic assessment of rehabilitative speech treatment in parkinsons disease. IEEE Trans. Neural Syst. Rehabil. Eng.22(1), 181190 (2014)

[9] El Moudden, I., et al. Feature selection and extraction for class prediction in dysphonia measures analysis: A case study on Parkinson's disease speech rehabilitation, Technology and health care: official journal of the European Society for Engineering and Medicine 25(4):1-16, April 2017

[10] Pandey, B., Pandey, D., An Integrated Algorithm for Dimension Reduction and Classification Applied to Microarray Data of

Neuromuscular Dystrophies, Indian Journal of Science and Technology, Vol 9(28), July 2016
[11] Saeys, Y., Inza, I., Larraaga, P., A review of feature selection techniques in bioinformatics, Bioinformatics, Volume 23, Issue 19, Pages 25072517, Oct. 2007
[12] Gibbons J. Nonparametric Statistical Inference. 5th Chapman Hall/CRC. 2010.
[13] MathWorks, sequentialfs, MATLAB documentation. [Online] https://www.mathworks.com/
[14] MathWorks, Feature Selection, MATLAB documentation. [Online] https://www.mathworks.com/
[15] https://www.quora.com/What-are-the-advantages-of-using-a-naive-Bayes-forclassification
[16] http://software.ucv.ro/cmihaescu/ro/teaching/AIR/docs/Lab4-NaiveBayes.pdf
[17] https://www.saedsayad.com/knearestneighbors:htmhttps :
[18] www:google:com=url?sa =trct=jq= esrc=ssource=webcd=19ved=2ahUKEwiSrKXy3anfAhUMxIsKHUGQDI0QFjASegQIABABurl =