Computers, Materials & Continua
DOI:10.32604/cmc.2022.020109
Article

Tech Science Press

Enhancing Parkinson's Disease Diagnosis Accuracy Through Speech Signal Algorithm Modeling

Omar M. El-Habbak[1], Abdelrahman M. Abdelalim[1], Nour H. Mohamed[1], Habiba M. Abd-Elaty[1], Mostafa A. Hammouda[1], Yasmeen Y. Mohamed[1], Mohanad A. Taifor[1] and Ali W. Mohamed[2,3,*]

[1]School of Information Technology and Computer Science, Nile University, Giza, 12677, Egypt
[2]Operations Research Department, Faculty of Graduate Studies for Statistical Research, Cairo University, Giza, 12613, Egypt
[3]Wireless Intelligent Networks Center (WINC), School of Engineering and Applied Sciences, Nile University, Giza, 12677, Egypt
[*]Corresponding Author: Ali W. Mohamed. Email: aliwagdy@gmail.com
Received: 09 May 2021; Accepted: 21 June 2021

Abstract: Parkinson's disease (PD), one of whose symptoms is dysphonia, is a prevalent neurodegenerative disease. The use of outdated diagnosis techniques, which yield inaccurate and unreliable results, continues to represent an obstacle in early-stage detection and diagnosis for clinical professionals in the medical field. To solve this issue, the study proposes using machine learning and deep learning models to analyze processed speech signals of patients' voice recordings. Datasets of these processed speech signals were obtained and experimented on by random forest and logistic regression classifiers. Results were highly successful, with 90% accuracy produced by the random forest classifier and 81.5% by the logistic regression classifier. Furthermore, a deep neural network was implemented to investigate if such variation in method could add to the findings. It proved to be effective, as the neural network yielded an accuracy of nearly 92%. Such results suggest that it is possible to accurately diagnose early-stage PD through merely testing patients' voices. This research calls for a revolutionary diagnostic approach in decision support systems, and is the first step in a market-wide implementation of healthcare software dedicated to the aid of clinicians in early diagnosis of PD.

Keywords: Early diagnosis; logistic regression; neural network; Parkinson's disease; random forest; speech signal processing algorithms

1 Introduction

1.1 Parkinson's Disease and Its Relationship with Speech Impairment

Parkinson's disease (PD) is a brain disorder whose symptoms include stiffness, shaking, uneven gait, and difficulty with walking and coordination [1]. It is one of the most common neurodegenerative disorders in the world and it affects people's lives at an epidemic rate. The cause behind PD is the impairment or death of dopaminergic neurons in the basal ganglia and the

substantia nigra. This leads to a dopamine deficiency, yielding the motor problems of Parkinson's. The full causes behind neural death are still unknown, so there is no definitive medical test that can identify the disease, making it ostensibly problematic to accurately diagnose [1].

Studies indicate that approximately 89% of PD patients experience speech and voice disorders, including a soft, monotone, and hoarse voice, coupled with hesitancy and uncertain articulation. This is because of the disordered motor system that comes in conjunction with PD. Poor muscle activation leads to bradykinesia and hypokinesia can carry on to the muscles involved in speech, possibly leading to reduced locomotion of the respiratory system, larynx, and deficient articulation [2]. An in-depth study specified that when the basal ganglia in the cerebral hemisphere of the brain is affected (as is apt to occur with PD) dysarthria might take place when muscle control is involved in the pronunciation mechanism, as displayed in Fig. 1 [3]. With the progression of PD, vocal folds occur—changes in the vocal cords. The vocal cord muscles become thinner and less taut, affecting their vibration performance and inducing the development of a gap in between the cords. The leakage of air through this gap is what results in the softness or hoarseness that is noticeable in the speech assessment of PD patients [4]. Furthermore, a closer look at the speech pattern identifies short bursts of words with pauses represented in longer, inappropriate silences before speaking again [5]. All of the previous findings identify exactly what to look for when trying to relate between speech signals and potential PD diagnosis.

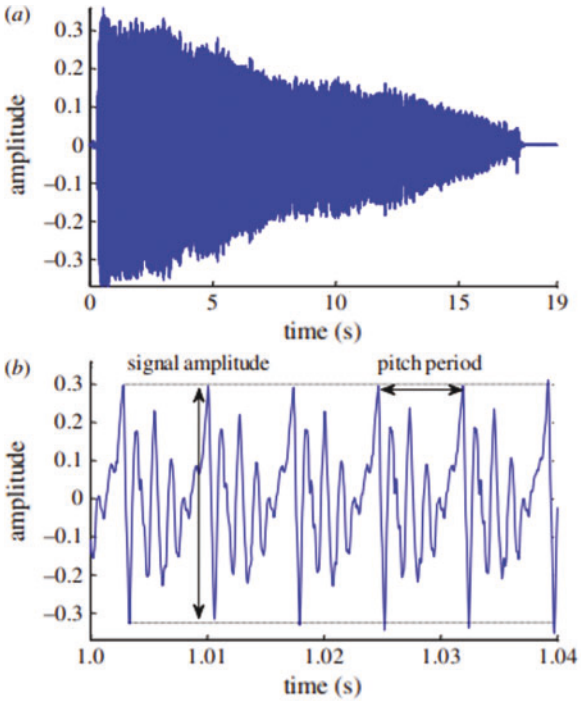

Figure 1: Graph (a) represents the signal sustained vowel while graph (b) shows the same signal magnified in the time axis. A close inspection of the signal in (b) shows it is not exactly periodic, a property of dysphonia [3]

1.2 Difficulty of Parkinson's Disease Early Diagnosis and Detection

It is very important to note that studies have shown that initial diagnoses conducted by general neurologists showed erroneous results in 24% to 35% of the cases upon postmortem patient examination [6]. To this day, diagnostic practices are mainly reliant upon clinical assessments. This is because there are no diagnostic biological markers for Parkinson's disease. Hence, in typical clinical settings, identifying cases of early-stage PD presents a diagnostic challenge. The reason this challenge is evident especially in the early stages of the disease is due to symptoms being highly subtle. Diagnosis relies on detecting the presence of three of the four cardinal motor signs (bradykinesia, tremor, rigidity, and postural instability); this criteria becomes inefficient because of the existence of a multitude of movement disorders sharing the same symptoms (vascular parkinsonism, essential tremor, multiple system atrophy) [7]. The situation can be elucidated by a well-thought conclusion made by Rizzo et al. on the accuracy of clinical diagnosis of PD: "The overall validity of clinical diagnosis of PD is not satisfying. The accuracy did not significantly improve in the last 25 years, particularly in the early stages of disease, where response to dopaminergic treatment is less defined and hallmarks of alternative diagnoses such as atypical parkinsonism may not have emerged" [8].

This paper uses machine learning and deep learning models to enhance the accuracy of early-stage PD diagnosis through classification of processed speech signals. Unlike previous research, which focuses mainly on the speech signal algorithm processing phase, this paper is dedicated to the pursuit of enhancing classification accuracy as much as possible. A gap in the research persists—maintaining the same ultra-specific scope of said algorithms while prioritizing the experimentation of both machine and deep learning methods on the resulting processed data. This voice-mixed-with-computer approach minimizes human error and provides a unique yet robust substitute for the outdated diagnosis techniques practiced in clinical settings worldwide.

Studies are showing that early intervention in PD could potentially help preserve neuron functionality, reduce symptoms, slow disease progression, and improve patient quality of life (QoL) [9]. Additionally, early diagnosis of PD is crucial because treatments such as levodopa/carbidopa are extensively more effective during early administration [10]. From an economic standpoint, estimated annual costs for PD in the US alone are approximately $11 billion, with $6.2 billion in direct costs [11]. Considering that the largest portion of the costs are spent in the later stages of PD, at which symptoms are most acute, then it would be unimaginably more cost-efficient to find a way to address the disease in its early stages rather than spend this huge excess on treating it ulteriorly, and the same policy is applied to the patient's QoL.

The structure of the paper is as follows: the first section presents the introduction, clarifying the challenge at hand and the paper's proposed solution. The first section also includes the structure of the paper which clarifies the content being discussed in each section of the paper. The second section is a background/literature review, which mainly focuses on researchers' previous efforts to solve the issue of inaccurate early diagnosis of PD. The scope starts out general and narrows down to the implementation of computational methods using processed speech signals. The third section represents the methodology, and basically consists of a detailed description of the model implementation. This includes, but is not limited to, model selection, feature selection, and the implementation of each specific machine learning algorithm. Graphs and statistical equations are used as supportive evidence. The fourth section shows the experimental results. This is followed by a comparative analysis and brief discussion to contextualize the value of the paper's contribution. Finally, the fifth section provides the conclusion of the research and future recommendations.

2 Background and Literature Review

2.1 Efforts to Improve Early Diagnosis of Parkinson's Disease

Efforts to improve accuracy of early-stage PD detection have been headlined by biological markers and advancements in neuropathological findings. Using the latter as the gold standard, studies have indeed increased accuracy and called for diagnostic biomarkers [12]. Other researchers have taken intersectional approaches, such as looking for symptoms and biomarkers in cerebral fluid, while also performing tissue imaging and biopsies, as many neurodegenerative diseases are a product of misfolded proteins [13]. In a separate study, a variety of premotor symptoms were identified, and unique approaches such as diminished olfactory functions and REM behavioral sleep disorders were used in attempt of early diagnosis, accompanied by other means of detection such as sonography, MRI, and exceedingly complex neuroimaging techniques. Once again, biological biomarkers such as protein panels, auto-antibody testing, and a 5-gene panel proved to be excellent diagnostic markers [11]. However, a computational and statistical-based approach could spare a lot of human and time resources being exerted in manually and biologically attempting to refine accuracy. Even better, an intersectional approach between all these methods could perhaps be the coup de grâce to end or significantly minimize inaccuracy of PD diagnosis once and for all.

Perhaps the progress most relevant to this paper in the efforts to increase PD early diagnosis accuracy (outside of speech signal processing) comes in a study by Mohskova et al., in which hand movements were obtained (via a motion sensor) to detect PD through machine learning methods. The kinematic parameters of subjects with PD and a PD control group were obtained via three motor tasks–finger tapping, pronation–supination of the hand, and opening–closing hand movements. Different classifiers were used and the key point determination was conducted using maximums and minimums finder algorithm in order to determine the binary disease status (PD or non PD) of each subject. The results were highly informative, displaying 95.3% in finger tapping accuracy, 90.6% for opening–closing hand movements, and 93.8% for pronation–supination [14].

2.2 Speech Signal Processing Algorithms

There is an intricate web of steps taken to convert analog sound signals phonated by patients into numbers that the model can analyze. Such is the process carried out in feature extraction, or "extracting features characterizing the underlying patterns of the speech signals using signal processing algorithms" [15]. Dysphonia, or malfunction in voice production, is measured by a series of stages: subjects are brought in to record several volume variations of their voice, and after initial filtration to remove any phonations prone to error (short recording, ensuing of coughing, etc.), thousands of signals of the sustained vowel "a" are processed. The next step is feature extraction, which is concerned with specifics regarding voice oscillatory motion; this has to do with the vocal fold previously explained in the section above. The vocal fold oscillation pattern (vocal fold opening and closure) is almost periodic in healthy voices, meaning that the intervals of time between two successive cycles are almost equal where the vocal folds are apart or in collision. These oscillation intervals are regarded by speech scientists as "pitch period" or "fundamental frequency." While in healthy voices the vocal folds collide and remain together for a certain segment of this cycle, dysphonia is identified by an "incomplete vocal fold closure," resulting in unnecessary breathiness and turbulent noise in the lung airflow. Therefore, those with voice disorders cannot vocalize steady phonations, and this is where speech signal processing algorithms come in; these algorithms take into account the aforementioned physiological conditions and quantify this inefficiency to prepare digital data ready for analysis that ultimately aids in clinical decisions. In speech jargon, these algorithms are called "dysphonia measures." Dysphonia

measures are implemented on the thousands of speech recordings obtained from the subjects, and there exists a multitude of software packages such as Praat [15]. Another notable feature extraction method is tunable Q-factor wavelet transform which in a study performed better or comparable to the most recent and developed techniques in PD classification [16].

2.3 The Use of Machine Learning and Statistical Methods on Speech Signal Processing Algorithms

An important step in noninvasive PD diagnostic decision support was taken in perhaps the most similar study to the current one; a wide spectrum of speech signal processing algorithms (dysphonia measures) were analyzed using two statistical classifiers: random forests and support vector machines. Patients were asked to vocalize sustained vowels, from which 132 different dysphonia measures were computed. The results were beyond state-of-the-art, with nearly 99% accuracy of classification of ten dysphonia features, proving that this suggested approach can complement existing algorithms in assisting classifiers in differentiation between control and PD patients [17]. However, a limitation in this study is that classification was conducted on only ten dysphonic features, while in reality the number of characteristics in speech signal processing is exponentially more.

A separate study took this idea a step further by implementing non-linear analysis of the range of speech signal processing algorithms on the standard clinical score that determines PD symptom severity (Unified Parkinson's Disease Rating Scale or UPDRS). Along with the normal set of tasks required of the patient, the study tested accuracy using self-administered speech tests. Selection algorithms were used to filter for the best subset, which was pumped into non-parametric regression and classification algorithms. The results were more accurate than clinicians' predictions, showing about 2 points' difference. This suggests the advancement of this technology to scale it up to large-scale clinical trials [16].

Mustaqeem et al. [18] provides a good example of using a neural network to identify patterns in the voice. The researchers formulated a speech emotion recognition system using a stacked network with dilated convolutional neural network features. This specific research is not related to diagnostics of a disease with voice symptoms, but it represents another step taken towards using voice intonations and fluctuations to infer the state of the human subject.

3 Methods

3.1 Model Selection

It is no secret that quality of results and model accuracy depend ultimately on two factors:

— Data quality
— Model selection (then fine tuning that model to optimal performance)

Therefore, choosing which model to work with is a decision that in no way can be taken lightly. In fact, a variety of factors go into such a question, and these factors played a substantive role in influencing which models were selected to work with in this research.

Factors affecting model selection:

- Size of training data: A large dataset such as the one present in this research is better suited for low bias/high variance algorithms such as decision trees, random forest, and K-nearest neighbor.

- Accuracy: There will always be a tradeoff between accuracy and interpretability of output, as is represented in Fig. 2. In this case, because the goal is to achieve the highest accuracy possible, then a flexible model is highly preferred.
- Speed/training time: Models with higher accuracy will usually require a higher training time, such as SVM and random forest, while models like logistic regression are quicker to implement.
- Linearity: Kernel SVM and random forest are preferred for non-linear data, while logistic regression and linear SVM are preferred for linear data.
- Number of features: Because this dataset has an extremely high number of features, dimension reduction is necessary before continuing to input the data to a classification model [19].

Based on the previous factors, and taking into consideration that a classification model is required to divide between positive detection and negative detection (0's and 1's), a model with the following specifications is required:

Classification model—High size of training data (low bias/high variance)—High accuracy—Linear or non-linear data (tested to see)—High number of features.

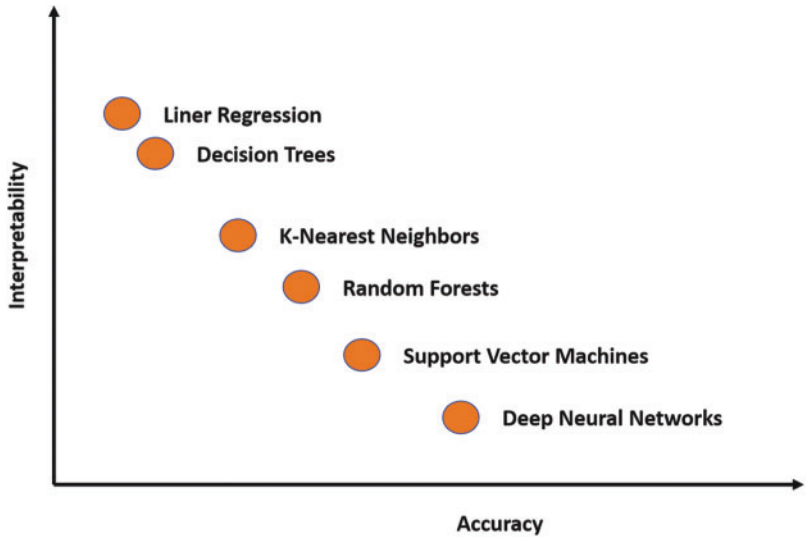

Figure 2: Accuracy—interpretability tradeoff

Thus, the decision was made to use random forest, logistic regression, and deep neural network algorithms. These three were chosen in specific because they covered all of the aforementioned criteria, but at the same time all three algorithms are significantly different than each other. Each algorithm is distant from the next on the accuracy-interpretability tradeoff spectrum. Each algorithm varies in run time and complexity of handled data. With such a unique challenge at hand, it is imperative to diversify the approach in order to identify the best point of attack in coming trials. The high versatility found between these 3 models facilitated algorithm experimentation, as through the results, it became evident which models were better suited for this unique task. Because of this, the findings hold significant importance for future research papers working on the same topic. Fig. 3 shows the framework of this study's methodology.

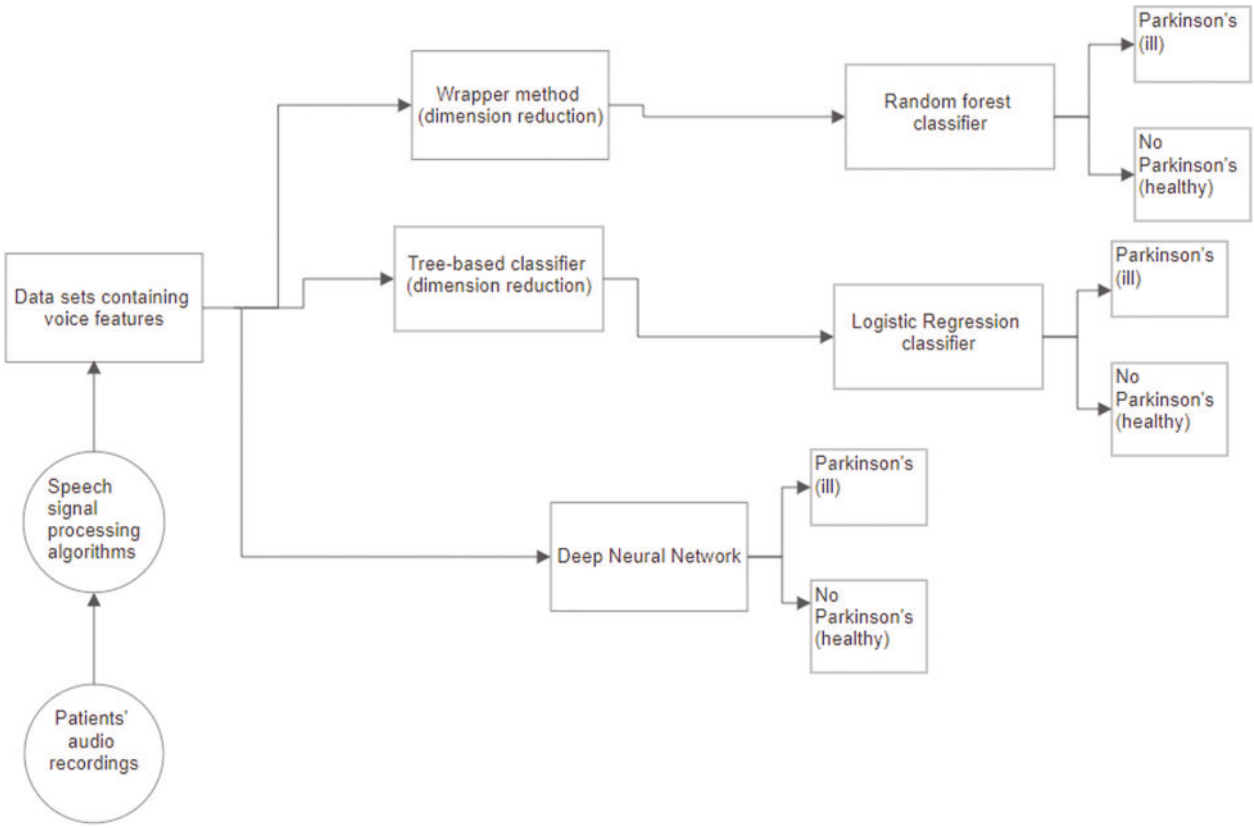

Figure 3: Block diagram of system framework [20]

3.2 Data Retrieval and Import

Data was retrieved as a csv file from a dataset on Kaggle™ (the Google online data science community), and this data was collected from UCI Machine Learning Repository [21].

The data was gathered from 188 patients (107 men, 81 women) with ages ranging from 33 to 87 (65.1Â ± 10.9), provided by the Department of Neurology, Faculty of Medicine, Istanbul University. The control group is made up of 64 people (23 men, 41 women) with ages ranging from 41 and 82 (61.1Â ± 8.9). The data collection process consisted of the subject sustaining phonation of the vowel "a" for three repetitions.

Attribute information: A variety of speech signal processing algorithms have been applied to the dataset, including Time Frequency Features, tunable Q-factor wavelet (TWQT), Wavelet Transform-based Features, and Vocal Fold Features in order to derive clinically significant information for PD diagnosis [22].

Data import: Using Anaconda Jupiter™ notebook in Python, a multitude of libraries were used, namely NumPy (for linear algebra and arithmetic equations), Pandas (data processing and csv file reading), and scikit-learn™ (free machine learning algorithm library for Python).

3.3 Data Visualization

Data visualization was done to identify general patterns within the data that would be difficult to recognize with just numbers. Various tools such as MatPlot library, Seaborn, and others facilitate this process. An example of this is a heatmap, such as the one included in Fig. 4, which uses Spearman's rank correlation coefficient to show the correlation between different features. It benefitted the research by indicating which features have high correlations and can be significant, therefore narrowing down the data extensively.

$$\rho = 1 - \frac{6\Sigma D^2}{n\left(n^2 - 1\right)} \tag{1}$$

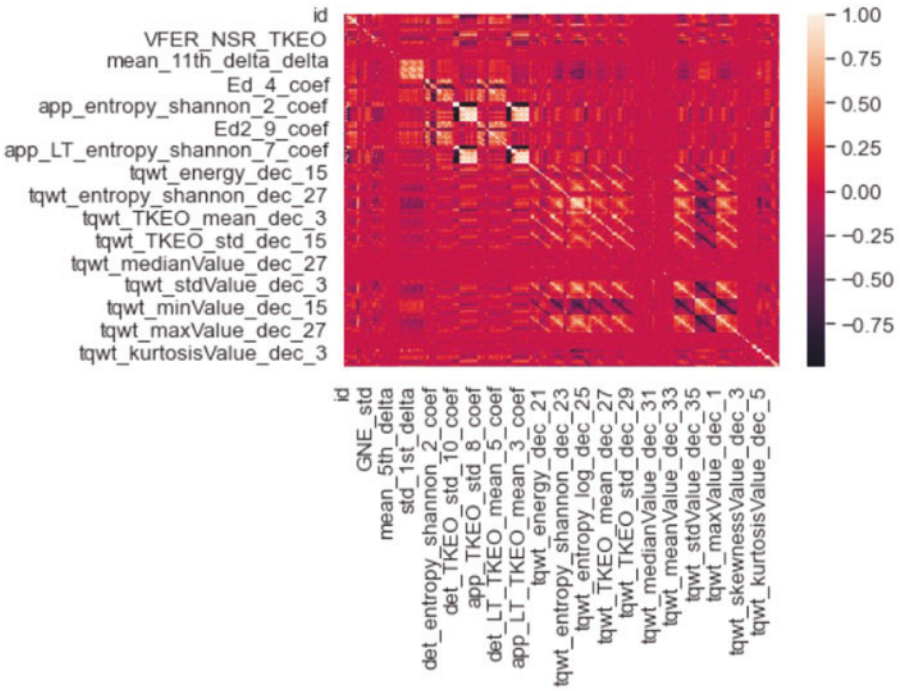

Figure 4: Heatmap of selected features

3.4 Feature Selection

One of the most crucial and significant phases of the data cleaning process, feature selection is the process in which the 754 features are slimmed down to only 15–20 via **dimension reduction**. This select batch of features is assumed to be the most significant group of features with the highest correlation, and therefore will yield the most useful data. Numerous dimension reduction methods can be used, but in this research the techniques used were wrapper method (for the random forest model) and tree-based classifier (for the logistic regression model). Each technique will be elaborated on in later segments.

3.5 Data Splitting

A commodity found in any predictive modelling, splitting is necessary to divide the dataset into a training set that the model can learn from and a test set that the model can test its

predictive accuracy upon. There are endless ways to split the data, but in this research the method used was 70/30. Fig. 5 gives an example of one of the ways used to split the data.

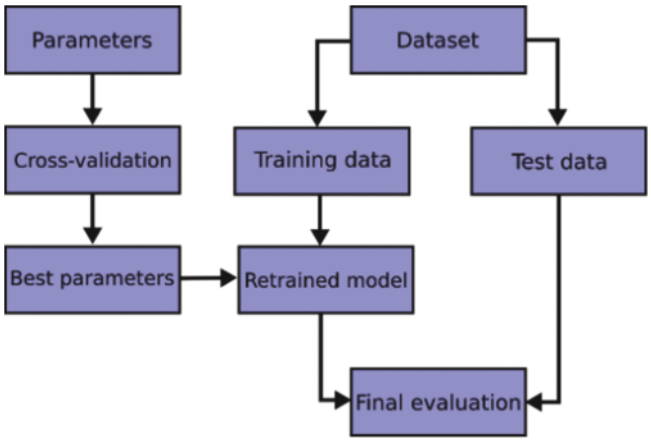

Figure 5: Data splitting diagram

3.6 Machine Learning Models

3.6.1 Random Forest Model

One of the most useful and accurate models, the random forest model is basically an aggregation of decision trees whose final decision is equivalent to the majority of final decisions of the trees composing it. After obtaining the dataset and importing it (as explained above), the dataset's first five rows is taken a look at using the head( ) function from the Pandas library.

Feature selection is done using the wrapper method, which follows a greedy search approach by evaluating all the possible combinations of features against the evaluation criterion. In this case for classification, the evaluation criterion is accuracy. The method then selects the combination of features that gives the optimal results for the algorithm [23]. ROC AUC is an integral part of this process (computing Area Under the Receiver Operating Characteristic Curve), in which the curve information is summarized in one number. ROC curve is the plot of the true positive rate against the false positive rate at all classification thresholds. It is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative. This is evident in Fig. 6.

$$TPR(T): T \to y(x), \quad FPR(T): T \to x$$

$$A = \int_{x=0}^{1} \mathrm{TPR}(\mathrm{FPR} - 1(\mathrm{x}))\mathrm{dx} \tag{2}$$

To test if the model works, train_pred and test_pred are compared. Parameters used to create the model:

n_estimators (number of trees to build before taking the maximum voting or averages of predictions)

random_state (facilitates replication of any solution)

max_depth (longest path between root node and leaf node)

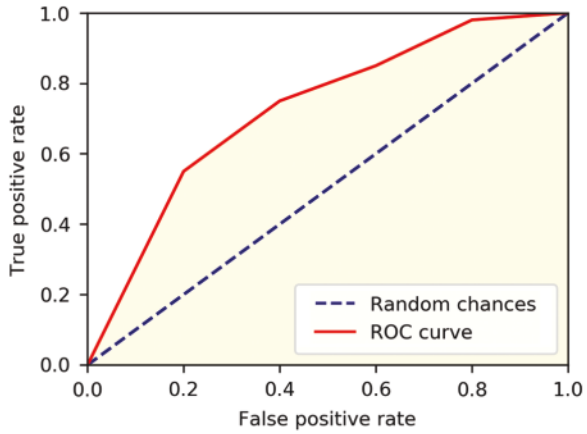

Figure 6: Graph of area under the ROC curve [24]

Finally, the data is fit to the random forest classifier and a confusion matrix is produced to indicate true positives, true negatives, false positives, and false negatives. Lastly, accuracy is produced.

3.6.2 Logistic Regression Model

One of the simpler models, logistic regression is a good way to test the data and see where it stands in terms of complexity and linearity. It is based on the standard sigmoid logistic function in statistics. Fig. 7 shows the graph of the standard sigmoid logistic function.

$$S(x) = \frac{1}{1 + e^{-x}} \tag{3}$$

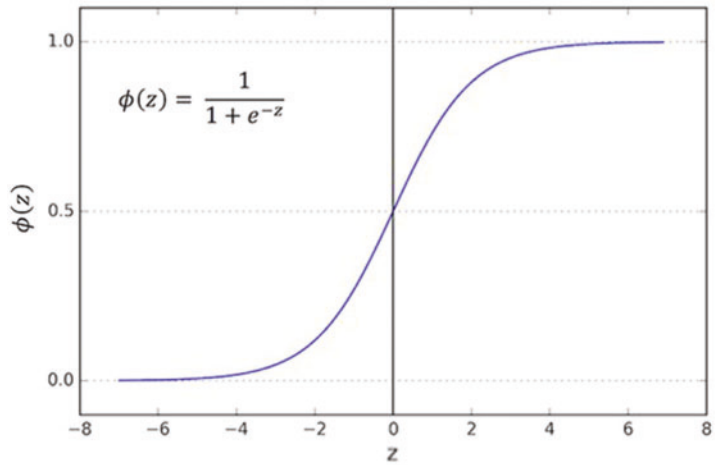

Figure 7: Graph of standard sigmoid logistic function

Steps:

— Import data and libraries (previously explained)
— Display data (previously explained)
— Select features using a tree-based classifier:

Because logistic regression is a restrictive model that depends on interpretability and time efficiency, using wrapper method would be too time-consuming. Similarly, univariate selection would be incompatible because it can only work with positive values, while the used dataset contains positive and negative values. Another technique, correlation matrix with heatmap, also requires a lot of time as well as high computational power. Therefore the method used, tree-based classifier, was the most suitable method, as it relies on decision trees. Although it might not yield the highest accuracy, it is much more efficient and works with negative numbers.

— Split data
— Model.

3.7 Deep Neural Network

In the context of this research paper, a neural network can be best defined as an intricate web of classifiers all linked together in the form of a network. This network contains input, output, and hidden layers that pick up on hard-to-detect patterns that simple classifiers would find difficulty with. Neural networks are extremely beneficial in situations where pattern complexity becomes a viable obstacle, and this became the driving force behind the idea of attempting to implement a deep neural network in this research.

Steps:

— Noteworthy libraries used here are TensorFlow, one of the main neural network frameworks, from which keras (API) and layers are imported. Pandas is also used to display the data entries in a table.
— Next, the data is split. The first split is 60% training, 40% validation. At the end of each epoch, the loss is evaluated as well as any model matrices of this data. Then the validation data is split into 80% training and 20% cross validation. Normalization of the data follows; training, valid, and test data are normalized. This is a transformation that maintains an output close to 0 and an output standard deviation close to 1.
— Subsequently, layers are created via a sequential model which trains stacks of layers. Data is input and output in the form of tensors, or 3D matrices. The type of layer used in this research is a regularly densely connected neural network layer (dense layer).
— Layers: 1 input layer, 7 hidden layers (ReLU activation function), and 1 output layer (Sigmoid activation function)
— Neurons: First five hidden layers: 70 neurons
Sixth hidden layer: 50 neurons
Seventh hidden layer: 30 neurons

Activation functions:

— ReLU (Rectified Linear Units): Allows use of non-zero threshold, change of max value of activation, and use of a non-zero multiple of input for values below threshold, as shown in Fig. 8.

$$R(x) = max\,(0, x) = \begin{cases} x & x > 0 \\ 0 & x \leq 0 \end{cases} \tag{4}$$

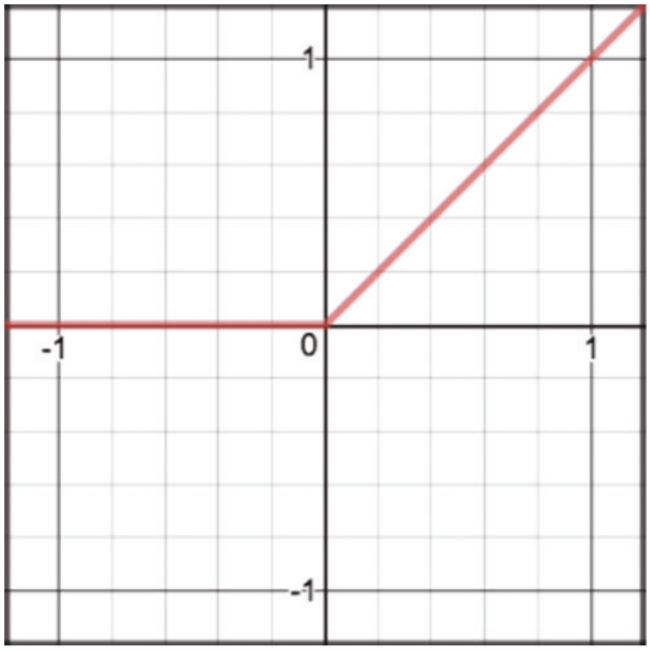

Figure 8: Graph of ReLU. Notice the threshold is greater than zero

— Sigmoid: Similar to ReLU but always returns value between 0 & 1 which is why it is used as the output. (see sigmoid function and graph in logistic regression section)

— After the creation of each layer, it undergoes batch normalization to provide more layer stability. In addition, the function Dropout() is used with parameter 0.1 to randomly remove 10% of the nodes, as this prevents overfitting.
— Optimization of the neural network is done using the notable method **"adam".**
— Another detail worth analyzing is the loss factor: The whole goal of training is to increase accuracy by removing losses. The loss factor used here is **cross entropy,** which is commonly used as a loss function when optimizing classification models. Fig. 9 indicates the relationship between cross entropy and predicted probability.

$$\boldsymbol{H}(\boldsymbol{p},\boldsymbol{q}) = -(\boldsymbol{y}\boldsymbol{log}(\boldsymbol{p}) + (\boldsymbol{1}-\boldsymbol{y})\boldsymbol{log}(\boldsymbol{1}-\boldsymbol{p})) \tag{5}$$

The callback used in this algorithm was **early stopping** of the training. Noteworthy parameters are:

— min_delta (determines the minimum change in the monitored quantity to qualify as an improvement)
— patience (counts how many epochs weren't improved on because the processing stopped)
— Batch size (how many samples are processed each time before the model is updated
— epoch (number of complete passes through the dataset before the improvement stops)
— Verbose (gives a status report of the training).

Lastly, accuracy is determined and visual representation is shown (elaborated on in the results section).

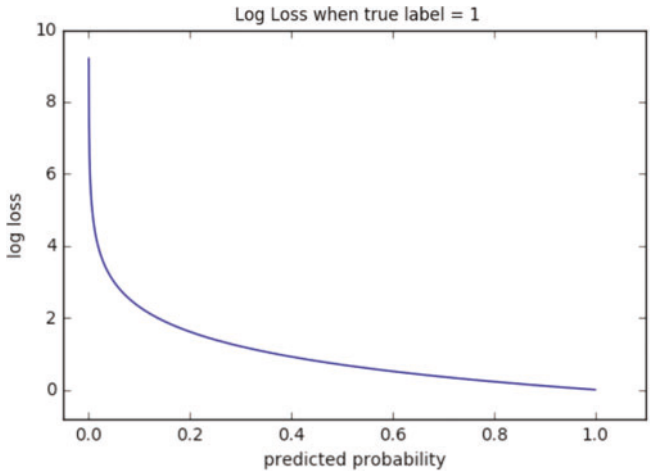

Figure 9: Relationship between cross entropy (log loss) and predicted probability

4 Results and Discussion

After applying different approaches to the same dataset in terms of feature selection and predictive modelling, the resulting accuracy shows discrepancies that can indicate significant differences in quality using all the experimented models, as the neural network showed once again its capability to increase accuracy with a highly pleasing result of nearly **92%** by detecting hidden patterns through its multiple layers, as depicted in Fig. 10. Additionally, random forest classifier model and the logistic regression model both yielded high accuracy percentages (**90.7%** and **81.5%** respectively). Looking at the confusion matrices, for the random forest: there were 32 true positives, and 174 true negatives, where true negatives are outcomes that actually were the same as the predictions by the model. There were a combined 21 false predictions, giving a laudable accuracy of 90%, which is highly efficient. For the logistic regression model, there was a combined 124 true predictions (12 true positives and 112 true negatives) and 28 combined false predictions, giving the model a commendable accuracy of 81.5%.

For the sake of a legitimate comparative analysis, the paper whose results were chosen to compare to this study's results was reference 16 "A comparative analysis of speech signal processing algorithms for Parkinson's disease classification and the use of the tunable Q-factor wavelet transform". This is because not only is that paper considered state-of-the-art (SOTA) in its domain, but it is also the paper most similar to the current research. Results of this study are included below in Fig. 11. It has an identical scope of study, yet the authors focus more on the part concerning speech signal processing algorithms, and less so on how to enhance the classifiers used in later steps of their research. The current study focuses on enhancing accuracy via experimentation on the machine learning algorithms. Not only that, but the current study proceeds also to use a deep neural network. Thanks to the above, the current study has yielded favorable results in classifier model accuracy over the SOTA study, as shown below:

The results in this figure show us that through different trials and by using various classification methods, accuracies varied from mid-60% all the way up to mid-80%. The highest accuracy found across all trials is 86%, when all feature subsets were used and classified by SVM (RBF). The highest accuracy obtained by logistic regression in the SOTA study is 85% (using all feature subsets). The current study achieved a logistic regression accuracy of 81.5%. Thus, the SOTA

study gains the edge in that regard. The highest accuracy obtained in the SOTA study using random forest is 85% (using all feature subsets). The current study achieved a random forest accuracy of 90.7%. The current study is superior in this regard. The current study went a step beyond anything done in the SOTA study by implementing a deep neural network in addition to the machine learning methods. The resulting accuracy was nearly 92%, almost 6% higher than any model in any trial conducted in the SOTA study.

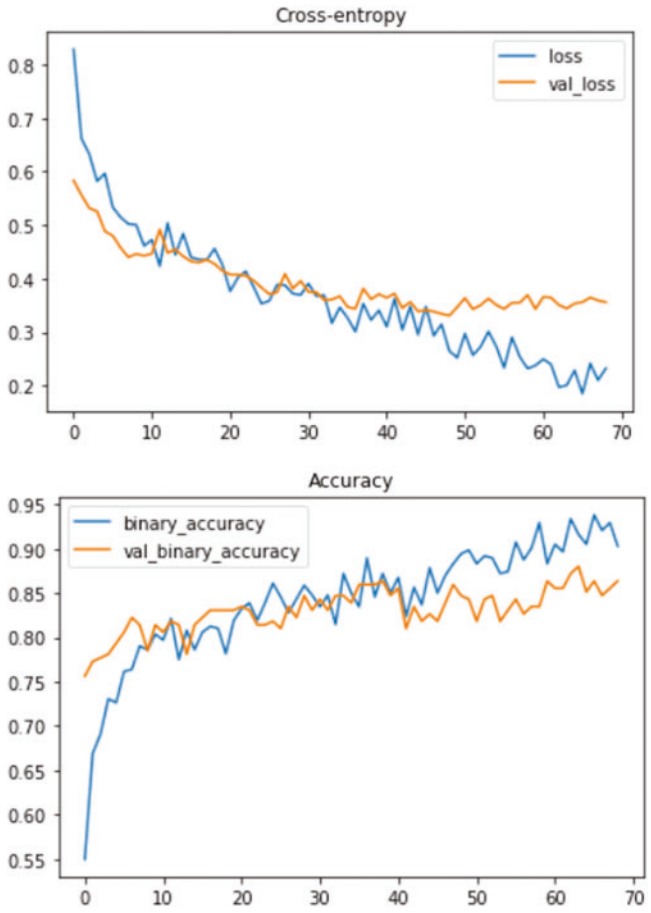

Figure 10: Graph depicting cross-entropy and accuracy of neural network

Results obtained with top-50 features selected by the mRMR filter on the combined feature subsets.

	All feature subsets except TQWT			All feature subsets except MFCC			All feature subsets		
	Accuracy	F1-Score	MCC	Accuracy	F1-Score	MCC	Accuracy	F1-Score	MCC
Naive Bayes	0.65	0.67	0.29	0.81	0.81	0.51	0.83	0.83	0.54
Logistic regression	0.81	0.79	0.45	0.83	0.82	0.51	0.85	0.84	0.57
k-NN	0.82	0.79	0.48	0.84	0.82	0.53	0.85	0.82	0.56
Multilayer perceptron	**0.83**	**0.81**	**0.50**	0.81	0.80	0.46	0.84	0.83	0.54
Random Forest	0.79	0.78	0.40	0.83	0.82	0.51	0.85	0.84	0.57
SVM (Linear)	0.81	0.80	0.46	0.84	0.83	0.54	0.83	0.82	0.52
SVM (RBF)	**0.83**	**0.81**	**0.50**	0.83	0.81	0.50	**0.86**	**0.84**	**0.59**
Average	0.79	0.77	0.43	0.83	0.82	0.51	0.84	0.83	0.55
Std. Dev.	0.07	0.05	0.08	0.01	0.01	0.03	0.01	0.01	0.02
Ensemble with voting	0.81	0.80	0.46	**0.85**	**0.84**	**0.57**	0.85	0.84	0.58
Ensemble with stacking	0.82	0.81	0.49	0.83	0.81	0.52	0.84	0.82	0.55

Figure 11: Results of state-of-the-art TQWT Parkinson's disease study [17]

Thus, the current study proves to have significantly enhanced the accuracy of the data present in its counterpart. This implies that if this research's methods had been used in the SOTA study, it would have actually yielded much more accurate results, although this wasn't the purpose of the SOTA as all the authors wanted to do was relatively compare variable features. However, in light of this numerical representation, the bigger-picture takeaway is evident: this study makes a significant contribution to the field of PD diagnostics, doing it in a unique and simple way—patients' voices.

5 Conclusion

The medical field can rest assured that the future of diagnostics is in good hands with the development of rapidly advancing machine learning technologies such as this one. Setting out with the goal of improving the early diagnosis of Parkinson's disease, this research has certainly achieved its task. After ending up with **three** different approaches (one deep neural network implementation, one machine learning model that is more flexible and aims for accuracy and non-linearity, and another machine learning model more restrictive, time-efficient, and linear) the researchers managed to conclude the study with a high prediction accuracy of Parkinson's Disease by modelling speech signal processing algorithms. With an accuracy of nearly 92% using the neural network, 90% using random forest and 81.5% using logistic regression, this proves as yet another step towards conquering diagnostic obstacles in the medical field, and is the beginning of a stable implementation of healthcare software in hospitals to aid clinicians with diagnostic decisions for PD patients. The limitations of this proposed method include: 1) the voice recordings of patients must be analyzed by speech signal processing algorithms as a preliminary step, in order to be broken down into features that the computational models can classify. Therefore, the data always has to be pre-processed. 2) Only 3 models were used in this research, which is quite a limited number considering the diversity that other extensive research papers show in model experimentation. Looking towards the future, the authors believe that the next step is to integrate speech signal processing with machine learning modelling. The authors also want to try more models to investigate differences in outcome. If these two caveats are taken care of, then this paper could very well be the first step on the road to worldwide implementation of a healthcare software that diagnoses PD by simply testing patients' voices for a matter of minutes.

Acknowledgement: We would like to recognize Dr. Ali Wagdy, Eng. Ruwaa Ibrahem, and Eng. Ahmed Kamal for their invaluable insight, constructive feedback, and undying support throughout the whole duration of this arduous research. They are our unsung heroes and the driving force behind the initiative to publish this paper.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] NIH National Institute on Aging website, "Parkinson's Disease," Retrieved on February 8, 2021. [Online]. Available: https://www.nia.nih.gov/health/parkinsons-disease.

[2] Parkinson's Foundation website, "Speech Therapy and Parkinson's," Retrieved on February 8, 2021. [Online]. Available: https://www.parkinson.org/pd-library/fact-sheets/Speech-Therapy.

[3] L. Yang, Y. He, X. Zou, J. Yu, M. Luo *et al.,* "Vocal changed in Parkinson's disease patients," *Archives of Biomedical and Clinical Research*, vol. 1, no. 1, pp. 1–2, 2019.

[4] Davis Phinney Foundation for Parkinson's website, "Closing the Gap for Voice Impairment in Parkinson's," Retrieved on February 8, 2021. [Online]. Available: https://davisphinneyfoundation.org/closing-the-gap-for-voice-impairment-in-parkinsons.

[5] The Voice Foundation website, "Neurological Voice Disorders," Retrieved on February 8, 2021. [Online]. Available: https://voicefoundation.org/health-science/voice-disorders/voice-disorders/voice-dysfunction-in-neurological-disorders/parkinsons-disease/.

[6] J. Jankovic, A. H. Rajput, M. P. McDermott and D. P. Perl, "The evolution of diagnosis in early parkinson disease," *Archives of Neurology*, vol. 57, no. 3, pp. 369–372, 2000.

[7] W. C. Koller and E. B. Montgomery, "Issues in the early diagnosis of Parkinson's disease," *Neurology*, vol. 49, no. 1, pp. S10–S25, 1997.

[8] G. Rizzo, M. Copetti, S. Arcuti, D. Martino, A. Fontana *et al.,* "Accuracy of clinical diagnosis of Parkinson disease: A systematic review and meta-analysis," *Neurology*, vol. 86, no. 6, pp. 566–576, 2016.

[9] D. L. Murman, "Early treatment of parkinson's disease: Opportunities for managed care," *American Journal of Managed Care*, vol. 18, no. 7, pp. S183–S188, 2012.

[10] ParkinsonsDisease.net website, "Diagnosis—Early Symptoms and Early Diagnosis," Retrieved on February 8, 2021. [Online]. Available: https://parkinsonsdisease.net/diagnosis/early-symptoms-signs/.

[11] F. L. Pagan, "Improving outcomes through early diagnosis of parkinson's disease," *American Journal of Managed Care*, vol. 18, no. 7, pp. S176–S182, 2012.

[12] C. H. Adler, T. G. Beach, J. G. Hentz, H. A. Shill, J. N. Caviness *et al.,* "Low clinical diagnostic accuracy of early vs advanced Parkinson disease: Clinicopathologic study," *Neurology*, vol. 83, no. 5, pp. 406–412, 2014.

[13] Parkinson's News Today website, "New Test Can Detect Parkinson's in Early Stages of the Disease," Retrieved on February 8, 2021. [Online]. Available: https://parkinsonsnewstoday.com/2018/02/14/new-test-can-detect-parkinsons-in-early-stages-of-the-disease/.

[14] A. Moshkova, A. Samorodov, N. Voinova, A. Volkov, E. Ivanova *et al.,* "Parkinson's disease detection by using machine learning algorithms and hand movement signal from leap motion sensor," in *2020 26th Conf. of Open Innovations Association (FRUCT)*, Russia, pp. 321–327, 2020.

[15] 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 Parkinson's disease symptom severity," *Journal of the Royal Society Interface*, vol. 8, no. 59, pp. 842–855, 2010.

[16] C. O. Sakar, G. Serbes, A. Gunduz, H. C. Tunc, H. Nizam *et al.,* "A comparative analysis of speech signal processing algorithms for Parkinson's disease classification and the use of the tunable Q-factor wavelet transform," *Applied Soft Computing*, vol. 74, no. 1, pp. 255–263, 2019.

[17] A. Tsanas, M. A. Little, P. E. McSharry, J. Spielman and L. O. Ramig, "Novel speech signal processing algorithms for high-accuracy classification of parkinson's disease," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 5, pp. 1264–1271, 2012.

[18] M. Mustaqeem and S. Kwon, "1D-CNN: Speech emotion recognition system using a stacked network with dilated CNN features," *Computers, Materials and Continua*, vol. 67, no. 3, pp. 4039–4059, 2021.

[19] KDnuggets website, "An easy guide to choose the right machine learning algorithm," Retrieved on February 8, 2021. [Online]. Available: https://www.kdnuggets.com/2020/05/guide-choose-right-machine-learning-algorithm.html.

[20] M. Mustaqeem and S. Kwon, "A CNN-assisted enhanced audio signal processing for speech emotion recognition," *Multidisciplinary Digital Publishing Institute*, vol. 20, no. 1, pp. 183–198, 2020.

[21] UCI Machine Learning Repository website, "Parkinson's Disease Classification Data Set," Retrieved on December 15, 2020. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Parkinson%27s+Disease+Classification#.

[22] K. Zhang, X. Liu, J. Shen, Z. Li, Y. Sang *et al.,* "Clinically applicable AI system for accurate diagnosis, quantitative measurements, and prognosis of COVID-19 pneumonia using computed tomography," *Cell*, vol. 181, no. 6, pp. 1423–1433, 2020.

[23] Analytics Vidhya website, "A comprehensive guide to feature selection using Wrapper methods in Python," Retrieved on February 8, 2021. [Online]. Available: https://www.analyticsvidhya.com/blog/2020/10/a-comprehensive-guide-to-feature-selection-using-wrapper-methods-in-python/.

[24] Towards Data Science website, "ROC Curve Transforms the Way We Look at a Classification Problem," Retrieved on June 8, 2021. [Online]. Available: https://towardsdatascience.com/a-simple-explanation-of-the-roc-curve-and-auc-64db32d75541.