An Explainable Machine Learning Model for the Prediction of Parkinson’s Disease using LIME on Speech Signals (Draft Version)

Richard Delwin Myloth Pavan Rajkumar Magesh Rijo Jackson Tom
Dept. of Computer Science and Engineering
CMR Institute of Technology
{rich17cs, pava17is, rijo.j}@cmrit.ac.in

Abstract

Speech impairments analysis has been used as an efficient tool for early detection of Parkinson’s disease (PD). In this paper we explore explainable models for the proposed Neural Network, SVM and Random Forest model developed to classify PD patients from the non-Pd patients utilising the speech samples. The performance of the method has been assessed with a reliable dataset from UCI repository. The proposed achieves an accuracy of 71.6%, 74.5% and 68.75% for the neural network, random forest and the SVM model respectively. Explainable models for the neural network and SVM have been generated with the help of LIME. The obtained results are compared with the results of the random forest model obtained by backtracking its decision path to obtain the most contributing features. The results are compared and its necessity are studied for explainable decision support in medicine.

Introduction

The World Health Organization (WHO) depicted neurological disorders as one of the major threats to public health. Parkinson’s disease (PD), stroke, multiple sclerosis, headache disorders, dementia, epilepsy, etc. are amongst the most common disorders. Parkinson’s disease (PD), initially called shaking palsy, first was described by James Parkinson in 1817[1]. Parkinson's disease is a progressive degenerative disorder of the basal ganglia that affects the initiation and execution of voluntary movements. It is the second most common neurodegenerative disorder after Alzheimer's disease.[2] The symptoms of Parkinson's usually begin gradually and get worse over time. Furthermore, the prevalence of PD is going to increase due to the aging population.

There is no definitive test for the diagnosis of PD, the disease must be diagnosed based on clinical criteria. Rest tremor, bradykinesia, rigidity, and loss of postural reflexes are generally considered the cardinal signs of PD[2][3]. Apart from these symptoms of the patients, currently, numerous physical tests such as (MRI, PET), etc are used to determine the plausibility of Parkinson's which are centred around assessing the dopamine levels in the brain. Although progressive parkinsonism which is referred to as Parkinson’s Disease, can be diagnosed upfront in patients with typical presentations of the above-mentioned cardinal signs, the differential diagnosis versus other forms of parkinsonism can be challenging, especially early in the disease when signs and symptoms of different forms of parkinsonism have greater overlap.[2]

Based on recent research findings, PD is much more than degeneration of the dopaminergic nigrostriatal system; the first neurons affected in PD are nondopaminergic. In addition to this population-based investigation have shown that no less than 15% of patients diagnosed with PD in the population do not satisfy strict clinical criteria for the disease, and roughly 20% of patients with Parkinson’s disease who have already received medical attention have not been diagnosed with the disease. Thus, by the time the disease has been diagnosed, around 60% of the nigrostriatal neurons degenerated, and 80% of striatal dopamine is depleted.[4] Apart from the four cardinal symptoms, speech disorders, visual hallucinations, etc. are few of the abnormalities reported according to [5]. Speech disorders in patients with PD are characterized by monotonous, soft, and breathy speech with variable rate and frequent word-finding difficulties. Collectively, these speech symptoms are called hypokinetic dysarthria and it affects 90% of its patients (of over 50 years of age).

According to [5] speech disorders associated with PD have been characterized by reduced voice volume with a tendency for it to decay over time, poor voice quality (dysphonia), reduced pitch inflection (hypoprosodia), range of articulatory movements (hypokinetic articulation), a tendency for speech articulation to festinate (rush), and hesitant and/or dysfluent speech. These disorders, collectively termed hypokinetic dysarthria (HKD) and the perceptual analysis of these symptoms are commonly used in clinical studies, however, such analysis is quite subjective to the practitioner and are influenced by various factors such as familiarity with the patients, their natural speech pattern and more. Thus, an approach devoid of human intervention might have fruitful effects in discerning the condition.

Machine Learning techniques, notably deep learning have proved to be reliable in the diagnosis process, due to which efficient and accurate classification of patients is possible. Even so, such techniques owing to its inherent nature, make it challenging to deduce the reasoning behind such classification, which undoubtedly is of utmost importance particularly in the healthcare industry. This paper explores not only the binary classification of PD and non-PD patients using neural networks, SVM, and random forest classification techniques but also aims to overcome the drawback of the uninterpretable nature neural networks and SVM by generating an explainable model with the help of LIME. LIME analyses the internal process to provide explainability to the classification. This provides an insight into interpreting the results on the possibility of PD. The outcome of the model is compared with that of a random forest model, which further clarifies the decisions.

In this paper, we have developed a random forest model, SVM, and a neural network model based on the dataset provided by. The results of the neural network model which is known for its un-interpretability –a black box- is then interpreted with the help of LIME (Local Interpretable Model-Agnostic Explanations). This interpretation is then compared to the outcome of the random forest model which is more explainable due to its architecture.

Literature Survey

The use of speech impairment analysis has been used extensively across several works for the early detection of Parkinson's Disease. One of the earliest works which used voice recordings of patients to diagnose PD was done by Ene M. et al. [5] in 2008 who proposed a probabilistic neural network (PNN) variant to discriminate between healthy people and people with Parkinson's disease. Indira R. et al. [6] proposed a back propagation-based approach with the help of artificial neural network. Boosting was used by filtering technique, and for data reduction principle component analysis was used. Caglar et al. [7] proposed similar ANNs, however using Multilayer Perceptron and Radial Basis Function Networks with linguistic hedges to discriminate between healthy people and people with PD achieving 96.7% accuracy and an impressive 100% specificity. A Agarwal et al. [8] on the other hand, predicted PD using speech signals using the relatively newer extreme machine learning methods achieving an accuracy of 81.55%.

Other techniques apart from neural networks were also used as seen in the work done by A.Tsanas et al. [9] who proposed a nonlinear signal approach to a large voice/speech dataset by applying wide range speech signal algorithm. This paper was performed using nonlinear regression and classification algorithm (SVM), and support visibility of frequent, remote, cost-effective, accurate UPDRS telemonitoring based on self-administered speech tests. A. Sharma et al. [10] also experimented with SVMs for classification and achieved accuracies of around 85%. Chen et al. [11] went further and a used nested-SVM classifier and came off with an improved accuracy of upto 93%.

Interpretability in healthcare still remains a primitive area of research as most studies focus on accurate diagnosis rather than proposing explanations for the outcomes. The models which are most interpretable would likely be that of tree and forest-based prediction models due to their ability to retrace the steps taken by the classifier. With reference to this, Sriram T et al. [12] proposed Random forest-based classification on a voice dataset achieving 90.26% accuracy. Similarly, R. Ramani et al. [13] proposed random tree-based PD classification on voice measurements achieving a claimed 100% accuracy. These models however do not depict true explainability. Ribeiro et al. [14] studied methods to interpret neural network classifiers on natural language data to provide explicit explanations and proposed the LIME framework. Any attempts at providing true explanations for PD classification models have not taken place and this paper attempts to be the first to do so with the help of the LIME framework on the speech signals dataset.

1. Materials and Method

1.1. Feature Engineering

The quality of a machine learning model is only as good as the quality of the features that constitute the dataset. The Feature Extraction of voice samples is integral to its development. Difficulty in speaking i.e. Dysphonia is characteristic of PD patients, as it affects their vocal cords and other facial muscles involved in speech. Subsequently, these features enable the differentiation of healthy individuals from the former. The Dysphonia measures were extracted from multiple types of sound recordings of sustained vowels, numbers, words, and short sentences[15]. The dataset used in this paper comprises of 26 linear and frequency-based features obtained from the voice samplings. They constitute 6 types of dysphonia parameters - frequency (jitter), pulse amplitude (shimmer), voicing, pitch harmonicity. Feature Extraction was carried out using the Praat acoustic analysis software by Sakar et al [15].

Feature	**Group**
Shimmer (dda) **Shimmer (local)** **Shimmer (apq3)** **Shimmer (apq11)** **Shimmer (apq5)** **Shimmer (local,dB)**	Amplitude Parameters
Number of pulses **Mean period** **Number of periods** **Standard deviation of period**	Pulse Parameters
Jitter (ddp) **Jitter (local)** **Jitter (rap)** **Jitter (local, absolute)** **Jitter (ppq5)**	Frequency Parameters
Number of voice breaks **Fraction of locally unvoiced frames** **Degree of voice breaks**	Voicing Parameters

Mean pitch	Pitch Parameters
Median pitch	
Standard Deviation	
Maximum pitch	
Minimum pitch	
Harmonic-to-Noise	Harmonicity Parameters
Noise-to-Harmonic	
Autocorrelation	

1.2. *Data acquisition*

The dataset used in this study is gathered from the open-source dataset made available on UCI Machine learning Repository compiled by Sakar et al. from Istanbul University.

The database comprises of training data belonging to 20 PD patients (6 females, 14 male) and 20 healthy individuals (10 females, 10 male) who appealed to the Department of Neurology in Cerrahpasa Faculty of Medicine, Istanbul University. The training data set consists of 1040 voice samples, out of which 40 random samples (one from each individual) have been separated for a self-consistency check of the developed predictor. An independent test set from Parkinson's disease people was also collected, under similar circumstances. It contains 168 recordings comprising of only the sustained vowels 'a' and 'o' recorded thrice each by 28 PD patients.

Apart from healthy individuals, the test group consists of PD patients who had been suffering from it for 0 to 6 years. The reliability of this dataset lies in the fact that the voice samples were selected by a group of neurologists from a set of speaking exercises that aim to lead to a more powerful sound of PWP. The feature vectors obtained were assigned with UPDRS values and the binary labels 0 and 1 for corresponding class information are denoted non-PD and PD individuals respectively by the neurologists.

1.3. *Data Pre-processing*

The Datasets were meticulously structured and organized by the researchers. Hence, it didn't require data cleaning. Due to the small size of the data set, no data reduction is essential either. The dataset isn't class imbalanced with over TK and TK. Owing to the ultimate aim of the paper to analyse the explainanbility of different features contributing to the results, dimensionality reduction is also eliminated. LDA, PCA and autoencoders may select features that are not interpretable [30].

Since this paper focuses on the binary classification of PD and non-PD patients, the UPDRS assessments were removed. The feature vector denoted as X, ultimately consists of 26 features as described in *section 1.1*. The class information labelled 0 and 1 is used as the target variable, y.

Scaling and Normalisation of X are implemented to achieve consistency in values across the feature vector and also assists in the swift convergence of the neural network model. Scaling and normalization were applied to the dataset before splitting it into training and test segments of the ratio 80:20.

Table 2: Dataset Split

Category	Healthy Control	Parkinson's Disease
Training	410	422
Test	110	98
Total	520	520

1.4 ML Models

Machine Learning has emerged as a reliable and preferred technique which is consistently adopted in developing various classification and regression models in health informatics. In this section, the following three different ML models implemented are discussed.

- Artificial neural network
- Random Forest
- SVM

The choice of selecting Machine Learning (ML) models is primarily reliant on the nature of the data. In this paper, we have implemented the techniques of neural networks (NN), support vector machines (SVM), and random forest (RF). NNs have the ability to learn and model non-linear and complex relationships which are vital in exposing any hidden relationships between the input and the outputs. It is especially beneficial considering the structure of the relationships exhibited in the dataset. After learning from the initial inputs and their relationships, the model can generalize and predict on unseen data by inferring relationships on unseen data. Additionally, many studies have shown that ANNs can better model heteroskedasticity i.e. data with high volatility and non-constant variance, given its ability to learn hidden relationships in the data without imposing any fixed relationships in the data. The RF model is well-known, to provide higher accuracy, can handle large data set with higher dimensionality. It can be used to rank the importance of variables in a classification problem naturally. In addition to this, it prevents overfitting trees in the model, if there are a significant number of them while also maintaining the interpretability of the model. SVM builds a hyperplane or set of hyperplanes in a manner that achieves maximum separation between the two classes. This optimization contributes to a model with a better generalization since the hyperplane is situated at equal and maximum distance from the classes. The hyperplanes are defined by a subset of training instances called, *support vectors*, which are used for predictions.

(i) Artificial neural network

Machine Learning has emerged as a reliable and preferred technique which is consistently adopted in developing various classification and regression models in health informatics. Deep learning, a technique in machine learning with its foundation in artificial neural networks, is emerging in recent years as a formidable tool in diagnosis and prognosis [17]. In this section, we discuss the outcomes on the application of the procured data to the neural network consisting of three hidden layers, one input and one output layer. The input weights and biases are arbitrarily chosen and fixed. Regularization function is employed to reduce overfitting.

The training data is fed into the sequential neural network using, a desired binary output $Y_i \in [0,1]$ is obtained for every input vector $X_i = (x_1, x_2, x_3 \dots x_n)$. The relationship between the input and the output can be represented by a function:

$$Y(X) = g(\sum_{j=1}^{H} w_{k,j}\, f\left(\sum_{i=1}^{n} x_i\, w_{j,i} + w_{j0}\right) + w_{k0})$$

Where H is the number of hidden neurons, $w_{i,j}$ represents the weight of the edge between the two vertices i and j, and w_0 represents the bias. The f represents is the activation function – Rectified Linear Unit (ReLU) used for the all the layers apart form the output layer which can be represented by the function r as :

$$f(x) = \begin{cases} 0, & x \leq 0 \\ x, & x > 0 \end{cases}$$

And g is the sigmoid function applied on the output layer:

$$g(x) = \frac{1}{1+e^{-x}}$$

The architecture of the neural network is depicted in figure 1 and it's summary is shown in Table.

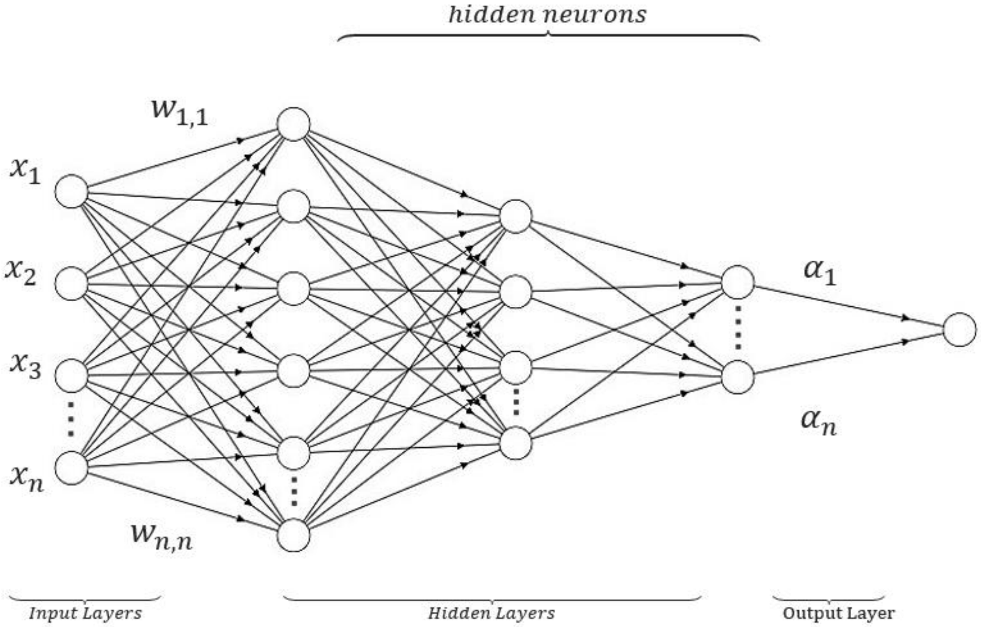


Figure 1

Layer (type)	Output Shape	Number of Parameter
dense_12 (Dense)	(None, 52)	1404
dense_13 (Dense)	(None, 26)	1378
dense_14 (Dense)	(None, 13)	351
dense_15 (Dense)	(None, 1)	14

Total params: 3,147

Trainable params: 3,147

Non-trainable params: 0

The neural network consists of three hidden layers, with 91 neurons (52, followed by 26 and then 13 neurons in the three hidden layers), an input layer(26 neurons) and 1 output layer(1 neuron). The activation functions of rectified Linear Units is used for all the hidden layers and sigmoid for the output layer.

Random Forest

Random forest is a machine learning model used for supervised learning for classification. It involves random sampling of training data points when building the trees and random subsets of features are considered when splitting nodes instead of averaging the prediction of trees. It is a transparent model, consisting of a collection of decision trees whose outcomes are aggregated into one final result. Decision tree hierarchical structures comprises of rules that do not alter data whatsoever, and preserves their readability. However, the combination of decision trees tends to lose it's transparent property. From an empirical study we have observed that the random forest model with 100 have better performance than the ones with 500, 600.

Cross Validation was applied to obtain the most efficient variation of the random forest model. Similar to ANN and SVM, a 10 fold cross validation was used. Empirically it was found that the version with a maximum depth of 70, minimum sample split of 20 and 400 estimators was found to be the best with a mean accuracy of 69.33% (+/- 2.25%). This model when trained on the entire dataset yielded 70.67% accuracy.

Cross Validation fold(k)	Accuracy
1	67.31%
2	73.08%
3	68.27%
4	72.12%
5	67.31%
6	69.23%
7	65.38%
8	71.15%
9	69.23%
10	70.19%

SVM

Similar to the procedure as the other two models with 10 fold cross validation it was observed that the classifier with radial basis function (RBF) kernel, cost value (c) parameter of 10 and kernel width (g) of 0.005 had a better performance yielding a mean accuracy of 64.62% (+/- 5.7). The final accuracy of SVC on the entire training dataset, which was validated on the unseen test data was found to be TK.

Cross Validation fold(k)	Accuracy
1	74.04%
2	63.46%

3	53.84%
4	65.38%
5	67.31%
6	66.35%
7	62.50%
8	69.23%
9	68.26%
10	55.77%

Threshold Tuning

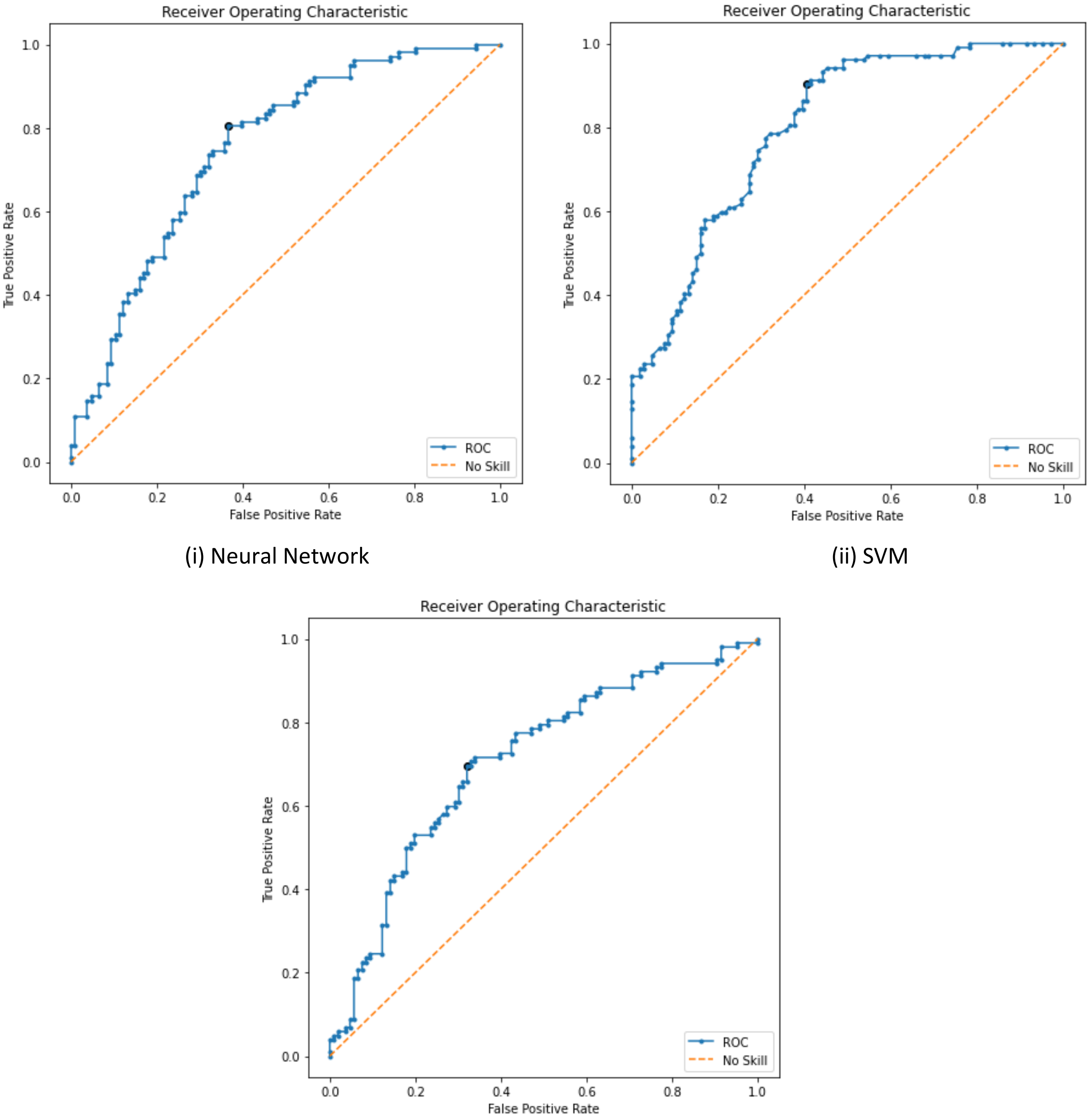
A threshold of 0.5 is the most commonly used threshold to classify the result into binary instances. However, it may not always be the optimal value which gives the best interpretation of the model. The best threshold value is obtained by plotting the Receiver Operating Characteristic (ROC) curve. The best threshold for each of the three models were determined by calculating the Geometric mean (G-mean) which is given by:

$$G_{mean} = \sqrt{Sensitivity * Specificity}$$

Where Sensitivity is the True Positive Rate (TPR =)and specificity is the True Negative Rate (TNR =).Higher values of G-mean corresponds to a better threshold value. In addition to this Youden’s J statistic also called the Youden index which captures the performance of a dichotomous diagnostic is computed which further verifies this, the Youden Index is given by:

$$Y = sensitivity + specificity - 1$$

The ROC cures plotted for Neural Network model (Figure TK.O), Random Forest (Figure TK.O) and SVM (Figure TK.O) is shown.



(i) Neural Network

(ii) SVM

(iii) Random Foerest

Figure TK ROC curves

	Threshold (Best)	TPR (Sensitivity)	FPR (Fall-out)	(LR+)	Youden index	G-mean			F-Score
Neural Network	0.5405					0.713			0.735
Random Forest	0.4400					0.732			0.780
SVM	0.4626					0.688			0.696

Table TK Metric

1.4 Evaluation Parameters

The effectiveness of the models has been gauged using the following parameters:

a) Average accuracy: It is the fraction of correctly determined cases (Healthy ad PD) to the total number of cases:

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

b) Specificity: It is the fraction of the PD cases correctly determined as PD.

$$Specificity = \frac{TN}{TN + FP}$$

c) Sensitivity: It is th fraction of healthy individuals correctly determined as healthy individuals.

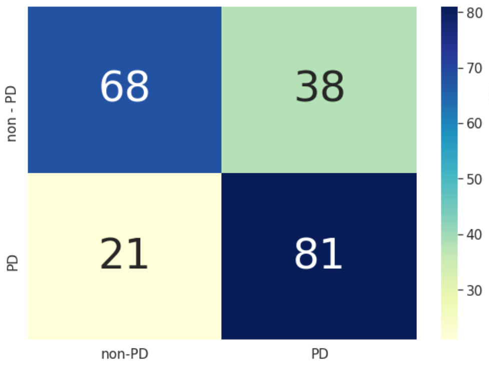
$$Sensitivity = \frac{TP}{TP + FN}$$

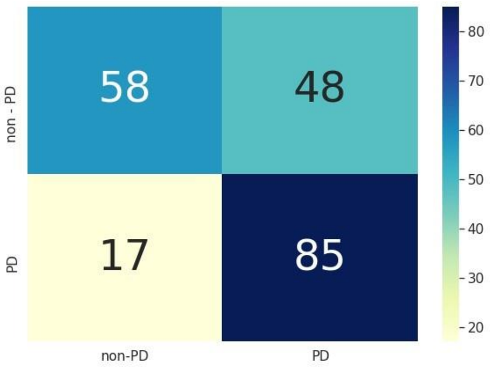
D) MCC: Matthew's correlation coefficient ranges from 0 to 1 where 1 corresponds to a perfect prediction whereas 0 for an exceedingly arbitrary prediction.

$$\text{MCC} = \frac{TP*TN-FP*FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \qquad (3)$$

True positive (TP) is the count of healthy patients predicted accurately as healthy, false positives(FP) is the count of diseased patients predicted as healthy, true negative (TN) is the count of diseased subjects accurately predicted diseased and false negative (FN) is the count of healthy patients predicted to be diseased.

1.4. Results





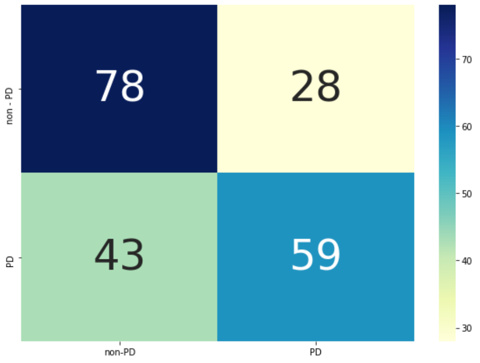
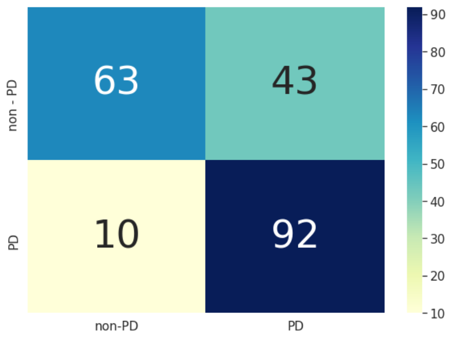
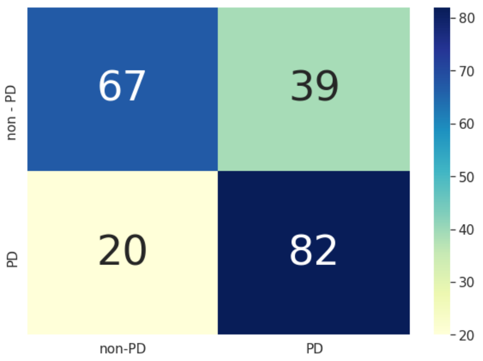


Figure 2.7.1: Neural Network *Figure 2.7.2: Random Forest* *Figure 2.7.3: SVM*

Confusion matrix of models before threshold tuning





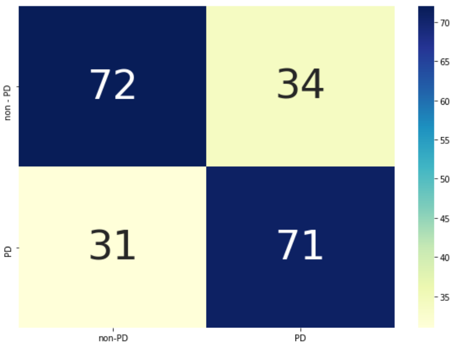


Figure 2.7.4: Neural Network *Figure 2.7.5: Random forest* *Figure 2.7.6: SVM*

Confusion Matrix after Threshold Tuning

Model	**Result**	**Overall Accuracy**	**Specificity**	**Sensitivity**	**MCC**
Neural Network	Average	69.71	79.41	60.37	0.4046
SVM	Average	62.21	71.86	53.11	0.2536
	Best	74.03	89.79	78.43	0.4837
Random Forest	Average	70.67	65.09	76.47	0.4179
	Best	73.07	79.59	79.24	0.4635

Table 3: Performance Results

Table 5 shows the average accuracy, specificity, sensitivity, and MCC for each model. The before and after values indicate the values before and after threshold tuning respectively. The confusion matrices are drawn to reflect the outcomes of the model after threshold tuning. The confusion matrices indicate that the number of false positives and false negatives across the models is relatively not low, which is further evident from the sensitivity and specificity of the models. Although individuals may be subjected to various other screening and tests for the diagnosis of Parkinson’s, this metric would raise concerns, where patients classified as false positives for PD may not get the vital timely treatments. Likewise, patients falling under false negatives may be subjected to unnecessary treatment. This serves as an instance which indicates the necessity for a model to be interpretable, where the decision of the model can be reasoned with.

2. Explainability

2.1. Need for Interpretability

Explainability is a characteristic of a model, denoting any action or procedure taken by a model with the intent of clarifying or detailing its internal function. It is of utmost significance in precision medicine, where experts require far more information from the model than a simple binary prediction for supporting their diagnosis. It is widely acknowledged as a crucial feature for the practical deployment of AI models.

All models are developed by humans and therefore there will be an element of human bias knowingly or unknowingly. Naturally, they exhibit biases which are inherent in the data itself. And often they lead to grave repercussions. A widely acknowledged example is the biased and unreliable decisions made by the AI in the COMPAS system which were used in Florida and other cities in the US. It’s a regression model which designed to predict the likelihood of a perpetrator to recidivate. It was found that the model predicted double the number of false positives for recidivism for African American ethnicities than for Caucasian ethnicities. This misjudgement is owed to multiple factors such as the choice of a single model, flawed dataset which was heavily biased to certain ethnicities. These systems expose the dire need of an interpretable system which is capable of making predictions accompanied with the parameters which led it to make such a judgement.

It is by virtue of these capabilities that AI methods are achieving unprecedented levels of performance when learning to solve increasingly complex problems by deducing relationships among multiple variables. Nonetheless, the limited explainablity of black-box Machine Learning (ML) models is a hindrance to its integration in real-world scenarios. Recent advancements to explainability of these black box models come under the research area of Explainable AI and include works such as DeepLIFT [14], RISE [15], SHAP [16] and LIME [17]. This study uses LIME to provide the explainability of the model created using DNN and SVM.

LIME (Local Interpretable Model-Agnostic Explanations)

LIME builds locally linear models around the predictions of an opaque model to explain it.

It is in essence a model-agnostic technique for post-hoc explainability designed to be incorporated into any model with the aim of extracting information of is internal functioning from its prediction. This technique attempts to understand the model by perturbing each feature individually drawing from a normal distribution with mean and standard deviation taken from the feature and correlating it with how the predictions change to derive its explainability.

LIME works by generating a new dataset consisting of permuted samples and the corresponding predictions of the black box model. On this new dataset LIME then trains an interpretable model, which is weighted by the proximity of the sampled instances to the instance of interest. In essence, LIME tests what happens to the when variations of the data is fed to the machine learning model. In the case of tabular data as this, LIME creates these new sample each feature individually, drawing from a normal distribution with mean and standard deviation taken from the feature.

The learned model should be a good approximation of the machine learning model predictions locally, but it does not have to be a good global approximation. This is also referred to as local fidelity. The output of LIME is a list of explanations, identifying the contribution of each feature to the prediction of a data sample. Its helps to draw insight into influence of feature changes on the prediction.

As stated in [] the local surrogate models can be expressed as

$$explanantion(x) = \arg min_{g\in G} \ (f, g, \pi_x) + \ \Omega(g)$$

The model g, which is a subset of the family of possible explanations G, is the explantion model of instance x that minimizes loss L. The loss measures how close the explanation is to the prediction of the original model f, while the model complexity Ω(g) is kept low. The model complexity is user determined and corresponds to the maximum number of features the model may use. The proximity measure π_x defines how large the neighborhood around instance x is that we consider for the explanation.

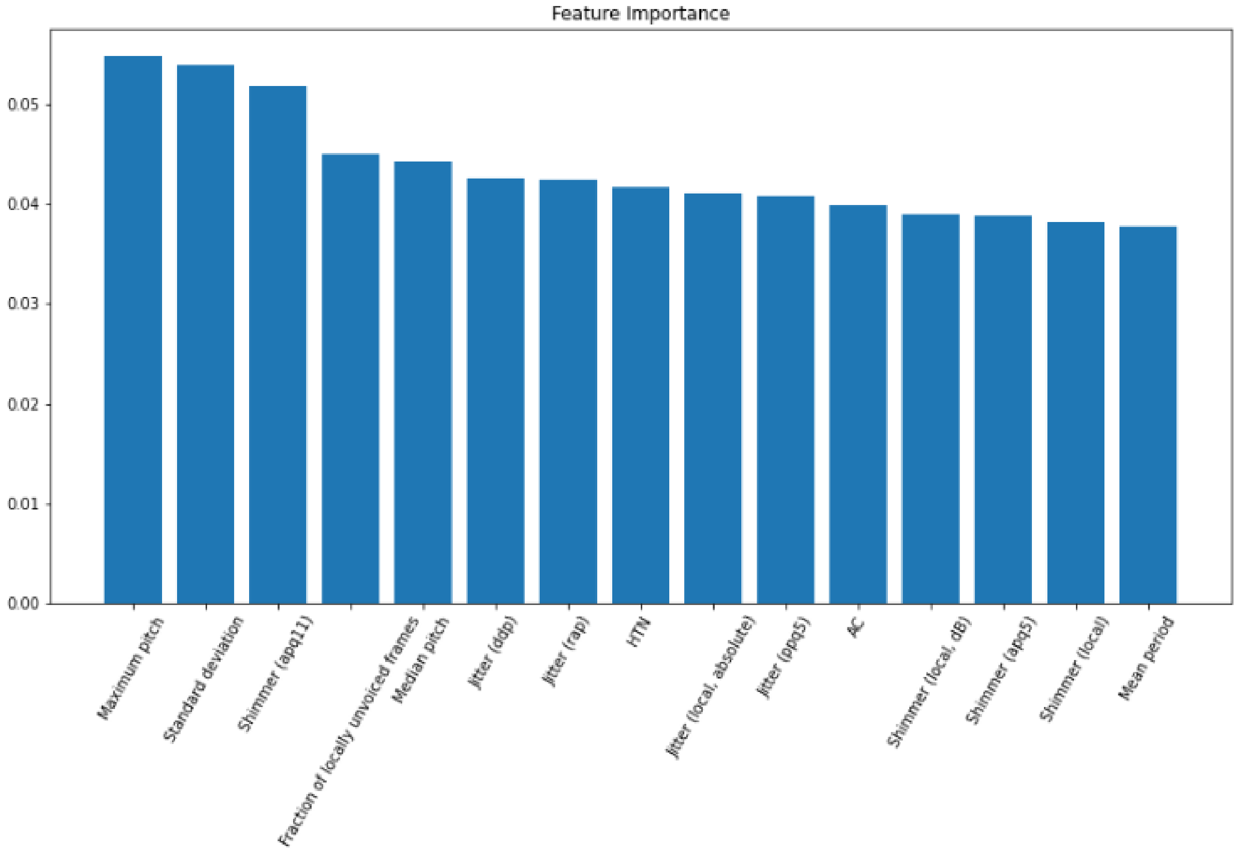
Implementation

The neural network and the SVM models had their explanation model generated, while setting the model complexity, $\Omega(g)$, to include all the features available in the dataset.

Results

Model	Neural Network (LIME)	SVM (LIME)	Random Forest (“feature_importance_”)
The most contributing features	i. Degree of voice breaks ii. Fraction of locally unvoiced frames iii. NTH iv. Maximum pitch	i. Minimum pitch ii. Fraction of locally unvoiced frames iii. Degree of voice breaks iv. Number of periods	i. Maximum pitch ii. Standard deviation iii. Shimmer (apq11) iv. Fraction of locally unvoiced frames

	v. Standard deviation vi. Standard deviation of period vii. Shimmer (apq11)	v. Shimmer (dda) vi. Shimmer (apq11) vii. Standard deviation of period	v. Median pitch vi. Jitter (ddp) vii. Jitter (rap)



The Table TK.O shows the most contributing features assessed through their weights for every instance of the test case which is correctly determined. The table shows the top seven features which appears the most among the aforementioned test cases. The results obtained for the neural network and the SVM classifier had been obtained with the help of the LIME framework. Whereas the result for that of Random Forest had been obtained with the help of "feature importance" parameter of the model provided by the scikit-learn framework.

From the models developed it was concluded that the Random forest yielded the best model with the highest accuracy of 71.63%, followed by the neural network model with an accuracy of 68.27%, and lastly the SVM with an accuracy of 62.61%. Thus, the features which contribute the most in each of the models are reflected in the results by their variation with few similarities. "Fraction of locally unvoiced frames" is observed to be a significant CF among the three models with it being the second most CF one in NN and SVM, and the fourth most one in RF. In addition, Shimmer (apq11) is present in all three models. The number common CFs between

- NN and SVM is 4
- NN and RF is 3
- SVM and RF is 2

It is also apparent that the NN and SVM share more features than they would individually with RF. This could be owed to the fact that NN and SVM have very similar fundamental architecture, which are naturally rather different from that of a RF. The number of common CFs between two models also reflect the accuracies of the models, the closer the accuracies of two models with each other the greater the number of common CFs. For instance, SVM and NN share more CFs than SVM and RF; Similarly NN and RF share more common features. This observation is not restricted to the above table alone, on examining the entire results the above assertion holds true as they do share the more CFs at the top.

References

[1] J. S. Almeida *et al.*, "Detecting Parkinson's disease with sustained phonation and speech signals using machine learning techniques," *Pattern Recognit. Lett.*, vol. 125, pp. 55–62, 2019, doi: https://doi.org/10.1016/j.patrec.2019.04.005.

[2] E. Tolosa, G. Wenning, and W. Poewe, "The diagnosis of Parkinson's disease," *Lancet Neurol.*, vol. 5, no. 1, pp. 75–86, 2006, doi: https://doi.org/10.1016/S1474-4422(05)70285-4.

[3] R. Mayeux *et al.*, "A Population-Based Investigation of Parkinson's Disease With and Without Dementia: Relationship to Age and Gender," *Arch. Neurol.*, vol. 49, no. 5, pp. 492–497, 1992.

[4] P. Gillivan-Murphy, P. Carding, and N. Miller, "Vocal tract characteristics in Parkinson's disease.," *Curr. Opin. Otolaryngol. Head Neck Surg.*, vol. 24, no. 3, pp. 175–182, Jun. 2016, doi: 10.1097/MOO.0000000000000252.

[5] M. Ene, "Neural network-based approach to discriminate healthy people from those with Parkinson's disease," *Ann. Univ. Craiova, Math. Comp. Sci. Ser*, vol. 35, no. january, pp. 112–116, 2008, [Online]. Available: http://inf.ucv.ro/~ami/index.php/ami/article/viewFile/250/245.

[6] I. Rustempasic and M. Can, "Diagnosis of Parkinson's disease using principal component analysis and boosting committee machines," *Southeast Eur. J. soft Comput.*, vol. 2, no. 1, 2013.

[7] M. F. Caglar, B. Cetisli, and I. B. Toprak, "Automatic recognition of Parkinson's disease from sustained phonation tests using ANN and adaptive neuro-fuzzy classifier," *J. Eng. Sci. Des.*, vol. 1, no. 2, pp. 59–64, 2010.

[8] A. Agarwal, S. Chandrayan, and S. S. Sahu, "Prediction of Parkinson's disease using speech signal with Extreme Learning Machine," in *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, 2016, pp. 3776–3779.

[9] 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 Trans. Biomed. Eng.*, vol. 59, no. 5, pp. 1264–1271, 2012.

[10] A. Sharma and R. N. Giri, "Automatic recognition of Parkinson's Disease via artificial neural network and support vector machine," *Int. J. Innov. Technol. Explor. Eng.*, vol. 4, no. 3, pp. 2278–3075, 2014.

[11] A. H. Chen, C. H. Lin, and C. H. Cheng, "New approaches to improve the performance of disease classification using nested--random forest and nested--support vector machine classifiers," *Res. Notes Inf. Sci. RNIS*, vol. 14, p. 102, 2013.

[12] T. V. S. Sriram, M. V. Rao, G. V. S. Narayana, D. Kaladhar, and T. P. R. Vital, “Intelligent Parkinson disease prediction using machine learning algorithms,” *Int. J. Eng. Innov. Technol.*, vol. 3, no. 3, pp. 1568–1572, 2013.

[13] R. G. Ramani, G. Sivagami, and S. G. Jacob, “Feature relevance analysis and classification of parkinson disease tele-monitoring data through data mining techniques,” *Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, vol. 2, no. 3, p. 2277, 2012.

[14] M. T. Ribeiro, S. Singh, and C. Guestrin, “‘ Why should I trust you?’ Explaining the predictions of any classifier,” in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[15] B. Sakar *et al.*, “Collection and Analysis of a Parkinson Speech Dataset With Multiple Types of Sound Recordings,” *Biomed. Heal. Informatics, IEEE J.*, vol. 17, pp. 828–834, 2013, doi: 10.1109/JBHI.2013.2245674.

[16] D. Ravi *et al.*, “Deep Learning for Health Informatics.,” *IEEE J. Biomed. Heal. informatics*, vol. 21, no. 1, pp. 4–21, Jan. 2017, doi: 10.1109/JBHI.2016.2636665.

[17] D. Ravì *et al.*, “Deep Learning for Health Informatics,” *IEEE J. Biomed. Heal. Informatics*, vol. 21, no. 1, pp. 4–21, Jan. 2017, doi: 10.1109/JBHI.2016.2636665.

Nevertheless, DNNs are a black box. Hence, its prediction is open to doubt. This uncertainty, naturally poses (some)? risk in sensitive use cases like that of in healthcare where people’s lives are at stake. Therefore, interpretability is of paramount importance especially in the healthcare sector. Thus, this paper will also discuss the interpretability and the technique adopted to achieve this interpretability in the later section.