Using Random Forest Classification to Identify Patients with Parkinson's Disease Based on Speech Characteristics

Camille Lake
Uniformed Services University of the Health Sciences
Bethesda, USA
camille.lake.ctr@usuhs.edu

***Abstract*- Parkinson's Disease is a neurodegenerative disease which affects more than 200,000 people in the United States alone each year. While there are some treatments, there is currently no cure. Recently, effort has been made to determine early signs of Parkinson's disease so that individuals may seek help and may mitigate the development of symptoms, increasing their chance of living longer and healthier. The purpose of this project was to use supervised machine learning to predict whether patients from the Parkinson's Disease cohort could be distinguished from healthy controls based on speech patterns. We found that, using a random forest classification model, we could predict the class that patients belonged to based on speech features with upwards of 75% accuracy. We hope that this model may be optimized in the future for diagnostic purposes, in an effort to discover patients with the early signs of Parkinson's Disease.**

***Index terms*- Parkinson's Disease, Random Forest Classification, speech patterns, Google Colaboratory.**

I. INTRODUCTION

Parkinson's Disease is a progressive neurological disease which typically affects older adults, affecting roughly 4% of the world's populations of adults over the age of 80. The disease is manifested by the degeneration of neurons over time, and is thought to be caused by the aggregation of Lewy bodies - proteinacious clumps of immunoreactive α-synuclein - which are resistant to the body's natural forms of waste disposal. The progression of Parkinson's Disease typically falls into distinct categories termed "Braak" categories 1-4. They are all characteristically defined by the degree of neurodegeneration accompanied by symptomology. Patients can exhibit a wide range of ages at which symptoms first occur, the degree of symptomology, and the speed at which the disease progresses, all of which likely have to do with the formation of toxic Lewy bodies and subsequent neurodegeneration. [1]

While the symptoms can vary from individual to individual, the general pattern of symptoms can include: body tremors, slow movement, rigid muscles, impaired posture and balance, loss of automatic movements, speed changes, and writing changes. [2]

One increasingly important factor which has been identified among the signs and symptoms of Parkinson's Disease is a change in speech patterns. One study conducted found that speech disorders are present in as many as 92% of patients with Parkinson's Disease [3]. Among these common speech patterns are abnormal voice modulation, hoarseness, decreased volume and a monotonous tone of voice [4]. Importantly, in a study by Robbins et. al., all of the patients studied had characteristic swallowing and speech patterns determined to be significantly different than the group of healthy controls. Worth noting is that the patient with Stage I Parkinson's Disease displayed 8 out of 14 swallowing abnormalities, demonstrating that changes to the throat and pharynx are not dependent on disease stage [4]. Further studies have corroborated these insights and added updated speech pathologies to this list. These studies highlight the significance of speech changes throughout the duration of disease.

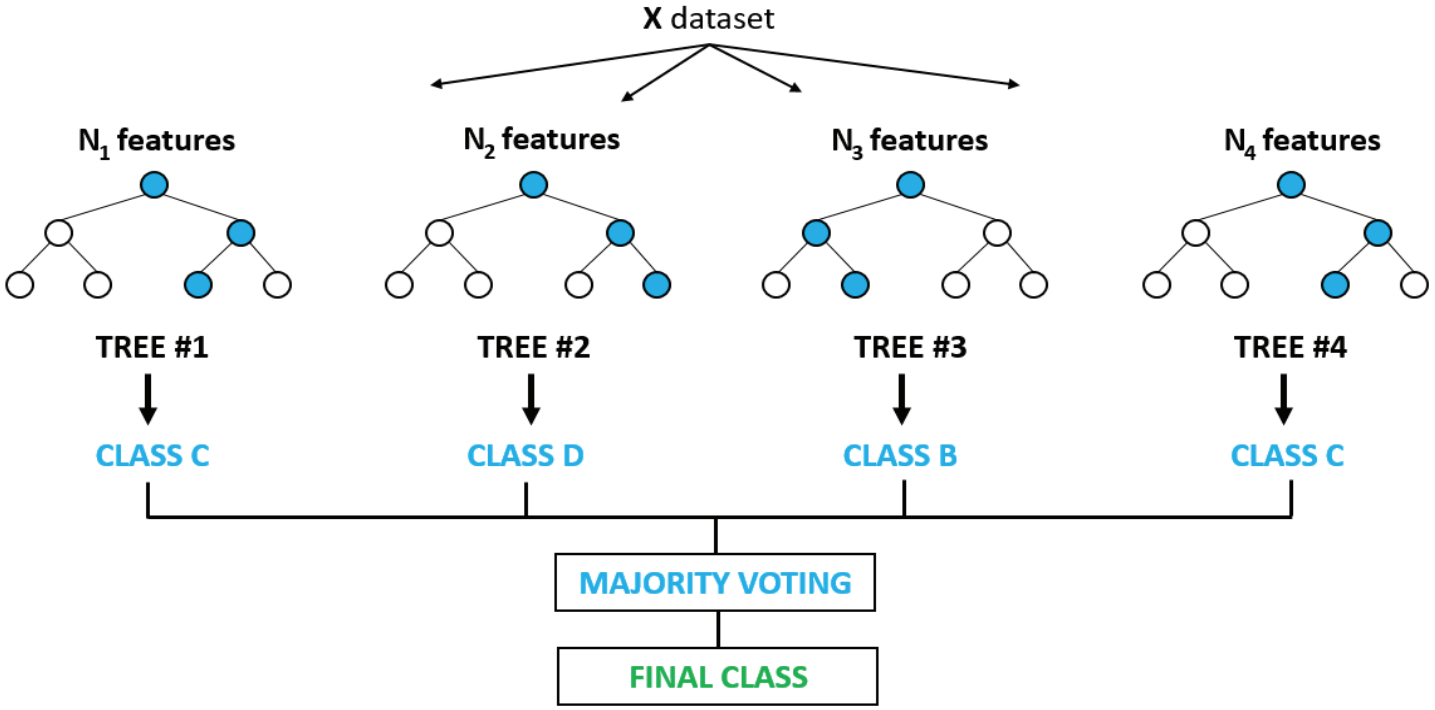
While there are treatments available for Parkinson's Disease, there are currently no cures for any diseases of this class - neurodegenerative synucleinopathies. A focus of late has been in disease identification in early stages in order to potentially delay the onset of symptoms for as long as possible. A potential avenue for identification of patients with the early signs of Parkinson's Disease relies on the distinct speech patterns aforementioned as well as others that have been identified. [5]

In this project, I sought to create a machine learning algorithm which would be able to identify patients with the early signs of Parkinson's Disease versus healthy controls based on speech patterns. My hope is that this tool can be optimized further (potentially by increasing the number of individuals used for training the algorithm) and become a routine part of diagnostics in the future.

II. RANDOM FOREST CLASSIFICATION

A. Model Concepts

The Random Forest Classification model is a supervised machine learning algorithm which is used to predict binary classifications based on a set of features in a dataset. According to the documentation for Scikit-Learn, random forest is a metaestimator which utilizes a multitude of individual decision tree classifiers on sub-samples of the data, using averaging to avoid over-fitting and improve accuracy [6].



As shown in Figure 1 [7], random forest classifiers aim to improve singular decision trees by optimizing both variance (the accuracy with which our model will predict the class in questions) and the bias (the ability of our algorithm to fit appropriately to the features within the dataset that are relevant to making an accurate prediction). They do so by utilizing multiple decision trees, taking an average vote from these trees, and determining the outcome based on this majority vote [7].

B. Model Parameters

The Random Forest Classifier utilizes a number of parameters, of which the relevant ones will be named (note, the following parameters are the ones used in this analysis and their descriptions are taken directly from the Scikit-Learn documentation) [6]:

- *n_estimators:* the number of decision trees in the forest
- *max_depth:* the maximum depth of the tree. In Sci-kit learn, the default allows nodes to be expanded until all the leaves are pure or until all of the leaves contain less than min_samples_split samples
- *random_state:* controls the randomness of the bootstrapping of the samples used when building trees
- *test_size:* used for determining the split between training and test sets (more applicable to the train_test_split object rather than the random forest classifier itself)

C. Justification for Using This Model

In general, this classification system was chosen on to answer this question because the dataset has many different features with a binary classification variable, of which random forest classification is ideal to solve. In addition, random forest classification offers a multitude of benefits.

The random forest classifier has proven itself to be a durable algorithm that is highly accurate across a wide variety of sample sizes. In addition, it is a significant improvement to a single decision tree as previously mentioned because of its optimization of variance, bias, and overfitting [7].

III. DATASET DESCRIPTION

The dataset in question was obtained from the original paper by Hlavnička et al. (2017) [8]. The dataset includes a multitude of features, of which only the features pertaining to differences in speech patterns were kept for final analysis. Though there are three unique classes which differentiate the cohorts, only two (patients with Parkinson's Disease (PD) and healthy controls (HC)) were kept for this analysis.

IV. CODE EXECUTION

This project was executed using Google's Colaboratory, but any platform which supports Python is acceptable for code execution.

A. Import All Necessary Packages

The packages necessary for this algorithm to run include: numpy, pandas, sklearn, StandardScaler, train_test_split, RandomForestClassifier,classification_report, confusion_matrix, accuracy_score, and GridSearchCV (optional).

B. Import Dataset

There are a multitude of ways to import datasets into Google Colaboratory; however, the following general pattern of import was followed.

```
from google.colab import drive
drive.mount('/content/gdrive')
import os
os.environ['KAGGLEC_CONFIG_DIR'] = "content/g-
drive/My Drive/kaggle_dataset"
%cd /content/gdrive/My Drive/kaggle_dataset
!kaggle datasets download -d ruslankl/early-bio-
markers-of-parkinsons-disease
!unzip early-biomarkers-of-parkinsons-disease
PD_df = pd.read_csv(r'dataset.csv')
```

This method of data import makes the following assumptions:
1. Person executing code has a kaggle account and has created a new API token
2. a google drive folder exists which is called“kaggle_dataset” and which the owner has placed the kaggle.json file into it

C. Data Manipulation

The dataset in question includes a number of features as well as observations which should not be included in the analysis. Overall, the source code was as follows to clean up the data:

```
PD_df_1 = PD_df.drop(PD_df.columns[1:41], axis=1)
PD_DF = PD_df_1.rename({' Participant  code
':'Participant_Type'}, axis=1)
PD_DF['class'] = PD_DF.Participant_Type.str[0]
target = PD_DF.pop('class')
PD_DF.insert(0, 'class', target)
PD_DF_new = PD_DF.drop(PD_DF.columns[1], axis=1)
PD_DF_new["class"] = PD_DF_new["class"].re-
place('P', 0)
PD_DF_new["class"] = PD_DF_new["class"].re-
place('H', 1)
PD_DF_new["class"] = PD_DF_new["class"].re-
place('R', 2)
col_names = list(PD_DF_new)
PD_DF_new['class']
col_names = list(PD_DF_new)
df_scaled = PD_DF_new.copy()
all_cols_no_class = df_scaled[col_names]
all_cols_no_class = PD_DF_new.loc[:, PD_DF_new.-
columns !='class']
scaler = StandardScaler().fit(all_cols_no_-
class.values)
scaled_all_cols_no_class = scaler.trans-
form(all_cols_no_class.values)
df_scaled = pd.DataFrame(scaled_all_cols_no_-
class)
df_scaled = pd.concat([df_scaled, PD_D-
F_new['class']], axis=1)
df_scaled_new = df_scaled.drop(df_scaled.loc[d-
f_scaled['class']==2].index)
df_new_1 = df_scaled_new.rename(col-
umns={'class':'24'})
```

In general, this code provided the following manipulation to the dataset: dropped the first 42 columns (the remaining columns are speech features); renamed the target variable to “class”; renamed the items in this variable to binary classifiers of 0,1,2; dropped the group containing the sleep disorder patients, and scaled the data using Standard_Scaler (except for the class variable). The new dataset is now ready for the random forest classifier to determine invididuals based on speech features.

D. Creating and Training the Random Forest Classifier

The following code was used to create the random forest classifier, split the data into training and test sets, and instantiate the new dataset with the random forest classifier object:

```
X = df_new_1.iloc[:, 0:23].values
y = df_new_1.iloc[:, 24].values
X_train, X_test, y_train, y_test =
train_test_split(X, y, test_size=0.5, ran-
dom_state=0)
classifier = RandomForestClassifier(max_depth=32,
n_estimators=100)
classifier.fit(X_train, y_train)
y_pred = classifier.predict(X_test)
```

This code allowed for the data to be split into the target and feature variable, further split into the training and test groups, and utilized to train the random forest classifier.

E. Accuracy Measurement

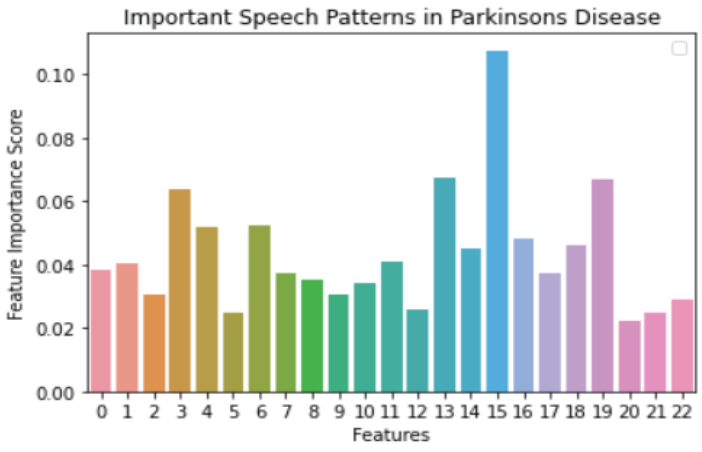
In order to determine whether the model was able to accurately predict individuals based on speech features, the following code was utilized:

```
print(accuracy_score(y_test, y_pred))
print(confusion_matrix(y_test, y_pred))
print(accuracy_score(y_pred, y_test))
```

The model in question displayed an accuracy of up to 75%, as displayed by the following:

```
0.75
[[ 9  4]
 [ 6 21]]
0.75
```

As shown in the figure below, features 15, 13, and 19 had the highest weight in increasing accuracy of the prediction. Of these three features: Acceleration of Speech Timing (15), Entropy of Speech Timing (13), and Duration of Unvoiced Stops (19), the first two were directly related to speech timing. Future studies should perhaps focus on understanding and utilizing this feature in building more accurate, powerful algorithms for diagnosis. [9]



V. SUMMARY AND CONCLUSIONS

In this project, a dataset with Parkinson's Disease patients versus healthy controls was assessed for the ability to train a random forest classifier algorithm to differentiate these groups based on speech patterns within the data. The random forest classifier was able to predict the patient class with an accuracy of upwards of 75%.

These results are important for a number of reasons: the first and most prominent reason being that it is now clear that machine learning can be used to differentiate between patients with particular symptoms in Parkinson's Disease; the underlying assumption being that there are detertable differences in speech patterns that algorithms can work through and differentiate. This is a big step in the direction of using machine learning as part of a comprehensive panel of diagnostics in the future. The second reason why this is important is because we now know that the model used specific features to best determine the differences; in particular, 2 of the 3 features with the highest weights were in the "timing" category of the features tested (as determined by the authors of the paper where the dataset originated) [9]. This suggests that the timing with which patients speak could be a key factor to focus on in terms of training future models.

VI. REFERENCES

1. Davie CA. A review of Parkinson's disease. Br Med Bull. 2008;86:109-27. doi: 10.1093/bmb/ldn013. Epub 2008 Apr 8. PMID: 18398010.
2. Mayo Clinic Staff, "Parkinson's Disease". DOA: 22 October 2020. https://www.mayoclinic.org/diseases-conditions/parkinsons-disease/symptoms-causes/syc-20376055
3. Martin WE, Loewenson RB, Resch JA, Baker AB. Parkinson's disease. Clinical analysis of 100 patients. Neurology. 1973 Aug;23(8):783-90. doi: 10.1212/wnl.23.8.783. PMID: 4578348.
4. Robbins, J.A., Logemann, J.A. and Kirshner, H.S. (1986), Swallowing and speech production in Parkinson's disease. Ann Neurol., 19: 283-287. doi:10.1002/ana.410190310
5. Huh, Y. E., Park, J., Suh, M. K., Lee, S. E., Kim, J., Jeong, Y., … Cho, J. W. (2015). Differences in early speech patterns between Parkinson variant of multiple system atrophy and Parkinson's disease. Brain and Language, 147, 14–20. doi:10.1016/j.bandl.2015.04.007
6. Scikit-Learn, "3.2.4.3.1. sklearn.ensemble.RandomForestClassifier". DOA: 22 October 2020. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html#:~:text=A%20random%20forest%20classifier.%20A%20random%20forest%20is,to%20improve%20the%20predictive%20accuracy%20and%20control%20over-fitting
7. Global Software Support, "Random Forest Classifier - Machine Learning". 23 February 2018, DOA: 22 October 2020. https://www.globalsoftwaresupport.com/random-forest-classifier/
8. Navlani, A. "Understanding Random Forests Classifiers in Python". 16 May 2018, DOA: 22 October 2020. https://www.datacamp.com/community/tutorials/random-forests-classifier-python
9. Hlavnička, J., Čmejla, R., Tykalová, T. et al. Automated analysis of connected speech reveals early biomarkers of Parkinson's disease in patients with rapid eye movement sleep behaviour disorder. Sci Rep 7, 12 (2017). https://doi.org/10.1038/s41598-017-00047-5