ASSIGNMENT COVER

Course code:	CIS6005
Course name:	Computational Intelligence
Assignment title:	Prediction and Classification of Parkinson’s Disease using Machine Learning
Instructor’s name:	Dr. K. Veropoulos
Student’s name:	Alexandros – Christoforos Mitronikas
Date:	20/12/2020
Word count	3843

DECLARATION

This work is the result of my own investigations, except where otherwise stated. This work has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

Signed Alexandros – Christoforos Mitronikas (Candidate)

Date 20/12/2020

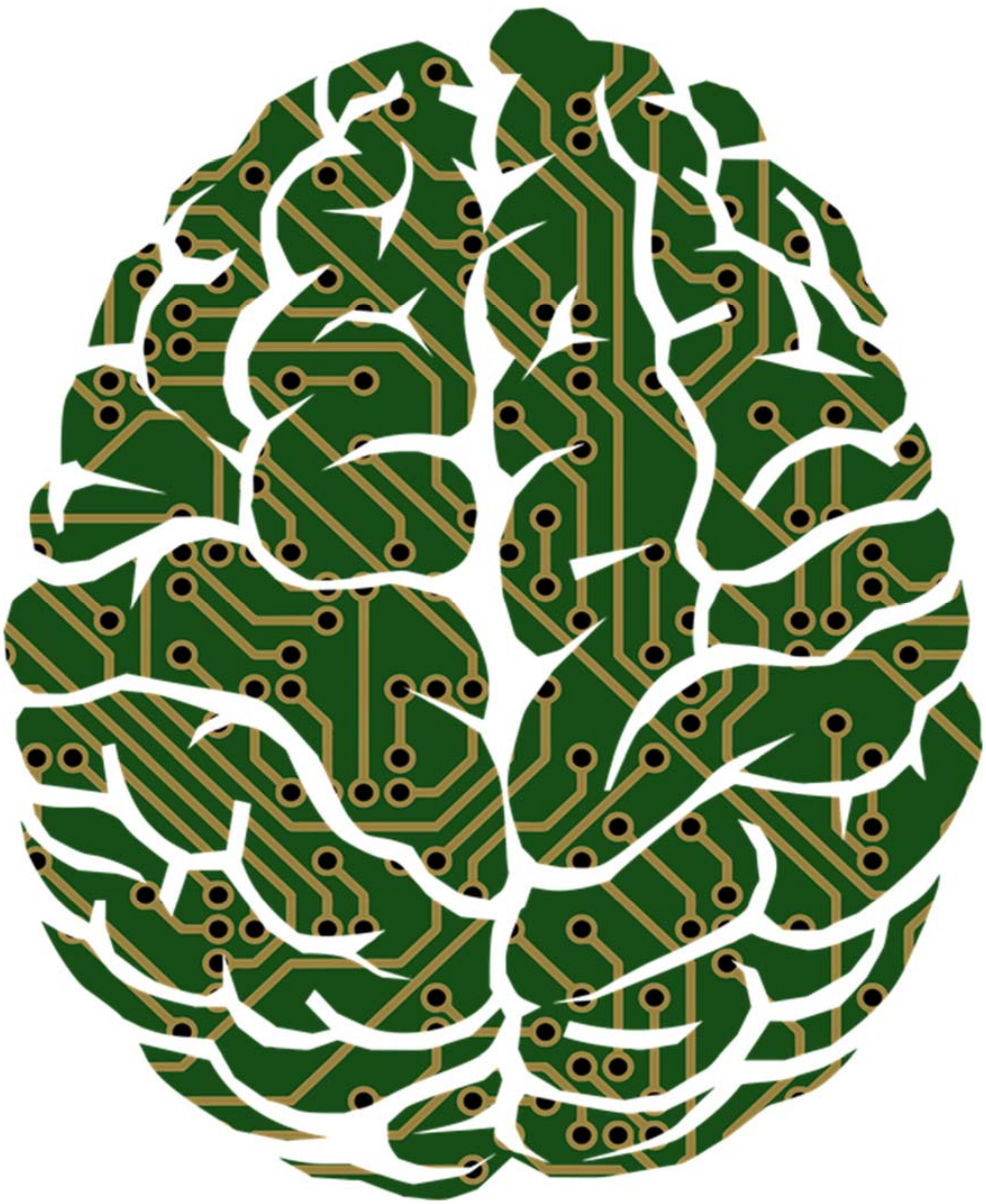
Table of Contents

ABSTRACT

This work has been conducted to explore through different approaches for the binary classification problem to individuals determining if they have, or do not have Parkinson's disease, using datasets with specialized attributes through machine learning algorithms. Viable models are viewed in addition with techniques of data analyzation in determining their best features to use for on-point results. Discussing on different methodologies, data manipulation, performance and results for better perceiving what a best approach on human speech datasets could look like for further implementation. The use of speech data for the recognition of Parkinson's disease is a feature that could be further explored, showing greater numbers with the help of AI compared to traditional costly and at times unreliable methods. The technology behind these algorithms can seem helpful to millions of people with its high accuracy saving their time, costs and even the need of transportation to a clinic using only speech samples.

Alexandros Christoforos Mitronikas
ST 20152336

PREDICTION AND CLASSIFICATION OF PARKINSON'S DISEASE USING MACHINE LEARNING



1. INTRODUCTION

The health sector has been advancing rapidly over the last decades with medicine and treatments at peak of attention in order for doctors to work better while improving patients' quality of life. Morosely, there are still numerous diseases not possibly cured yet. Such disease is the Parkinson's Disease which at early stages correct medication may show significant improvements on a patient's symptoms but without terminating it fully (McDermott, 2019). Moreover, there is not a fully accurate or reliable way to determine whether a person has Parkinson's disease. Using current technologies and tools such as statistical and machine learning we can recognize patterns from big volume of data quickly and reliably choose which attributes could lead us to likely results.

About 90% of clinically confirmed Parkinson's Disease cases are idiopathic (Physiopedia, 2020), recommending the extent of demand for research in order to define more accurate and exact diagnosis. It is a neurological disorder giving hard time financially for the patients in terms of treatment and initial diagnosis costs. Diagnoses for widespread sicknesses, as for example PD, do not depend on standard biochemical information for determining if a patient has or does not have the neurodegenerative disease. Clinical experts rely entirely upon actual tests visually examined based on impaired movement and hand-eye coordination practices for prior detection. At its origin, PD is caused by dopaminergic neurons in the substantia nigra pars compacta (Triarhou, 2013). The reduction in dopamine production extends the effects towards human cognition and ultimately leads to vocal and physical impairment

Many ways of approach for testing potential patients of Parkinson's disease are there but none has been definite of the diagnosis and evaluating its progress. Examination is subjective and based on each patient's symptoms, physical identification and medical history. Alternatively, imaging tests such as the MRI (Magnetic Resonance Imaging), SPET (Single Photon Emission Computed Tomography), PET (Positron Emission Tomography) and brain ultrasound could also be used for diagnosis but still requires a professional's contact and expose patients to radiation (Anonymous, 2019).

Speech disorder is one of the secondary but critical symptoms in patients with PD, shown as mumbled, slurred or slow speech, with voice becoming breathy, monotone and hoarse, lacking some of the usual volume changes. Using speech from patients for identifying Parkinson's disease is a low cost and an easy noninvasive task to even use for self-diagnostics. Speech impairments are symptoms to be met up to five years ahead of clinical diagnosis and starts degrading along with the disease's progression. It is for the best to consider speech-based Parkinson's disease identification and progress tracking as a tool for even remotely monitorization and diagnostics.

2. DATA CLARIFICATION

In recent years, a substantial amount of research committed for the connection between speech impairment and Parkinson's disease are being implemented examining various dysphonia measures from sustained vowels. Basically, gathered voice samples with additional personal information are critically analysed by distinguishing different attributes of dysphonia. Features such as the average, maximum or minimum vocal fundamental frequency are obtained from patient's voice records differing healthy subjects and PD patients. Data attributes may range based on sex, emotional state and age adding further critical information. Other data is determined from the shimmer, jitter levels, noise to harmonics ratio and vice versa (Senturk, 2020).

Table 1. Data features and descriptions (Senturk, 2020)

No.	Feature	Description
1	F_0 (Hz)	Average vocal fundamental frequency
2	F_{hi} (Hz)	Maximum vocal fundamental frequency
3	F_{lo} (Hz)	Minimum vocal fundamental frequency
4	Jitter (%)	Jitter in percentage
5	Jitter (Abs)	Absolute jitter in microseconds
6	Jitter: RAP	Relative average perturbation
7	Jitter: PPQ	5-point period perturbation quotient
8	Jitter: DDP	Difference of differences of periods
9	Shimmer	Local shimmer
10	Shimmer (dB)	Local shimmer in decibels
11	Shimmer: APQ3	3-point amplitude perturbation quotient
12	Shimmer: APQ5	5-point amplitude perturbation quotient
13	Shimmer: APQ11	11-point amplitude perturbation quotient
14	Shimmer: DDA	Difference of differences of amplitudes
15	NHR	Noise-to-harmonics ratio
16	HNR	Harmonics-to-noise ratio
17	RPDE	Recurrence period density entropy
18	D2	Correlation dimension
19	DFA	Detrended fluctuation analysis
20	Spread1	Nonlinear measure of fundamental frequency
21	Spread2	Nonlinear measure of fundamental frequency
22	PPE	Pitch period entropy

Many cases are there where missing or mistyped data are found in these vast and high complex information structures. There are plentiful techniques for the best organization of data each best fitting specific types, depending on their nature and complexity level. Some of these data mining methods are good concepts for approach while acknowledgement of human error situations takes place. Incomplete, noisy, inconsistent or duplicated data can be met in the real world; thus a series of pre-processes should take place to avoid misleading conclusions. This is the process of sorting and categorizing data into various types, forms or other classes that distinct them from each other. Methods other than classification are; association, clustering analysis, prediction, decision trees, neural network and more.

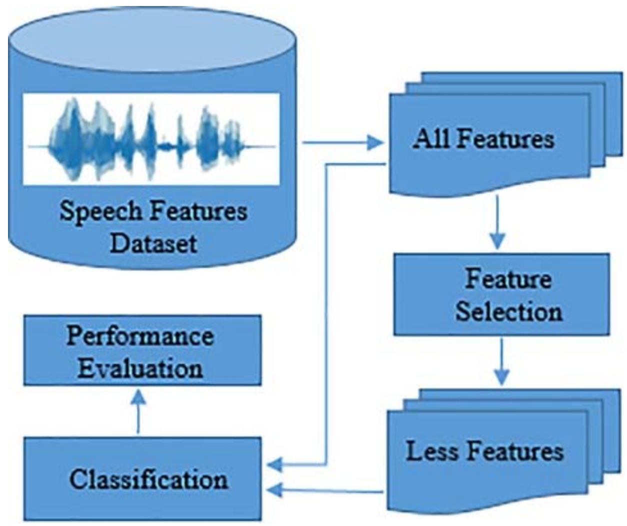


Figure 1. Speech data flow concept

3. OTHER AUTHORS

In this paper other authors' approaches working for the cause of PD prediction, identification and progress of potential patients will be analysed and discussed investigating within their methodology, data manipulation, performance and results. Within these researches, authors share similar thoughts of data records using mainly voice datasets but also other characteristic data, collected straight from patients or from online dataset sources. This does not stop them from using distinct methodologies, different types of machine learning algorithms or software choices. All factors will be further analysed specifically for each author in the next segments.

3.1 Analysing an approach

The first research discussed is Shounak Ray's journal article who at age of 16 in 2019 completed for the cause of Parkinson's disease binary classification problem (Ray, 2019). Looking for new results for better determining a potential patient's diagnosis through machine learning, he used two datasets including the human demographic, movement and speech data. Conducting statistical testing on each of the datasets independently, more than 30 different machine learning models are to be analysed through visualizations, analysis and testing. Suggesting that such a machine learning system based on speech, demographic and movement data concludes to a more accurate, cost-effective, easy and time-efficient process, determinising if an individual is healthy or has Parkinson's disease.

As mentioned above, different datasets were to be used hence the project was divided into three stages, each for distinct procedures in the investigation. For all two datasets, the patients were laminated across whether they had Parkinson's disease, early or late stage diagnosis also accounting their geographical location. All locations were fairly balanced in order to sustain valid results throughout North America, South America, West Africa and India.

A complete list of Demographic, Movement and human speech features follow to indicate the names of the attributes. For human speech the attributes have been categorized while there are many sub-features used in the confirmatory procedures that can be found in table 1 within the data clarification section.

Table 2. Author's datasets attributes (Ray, 2019)

Human Speech (n = 4485)	**Demographic-Movement (n = 3700)**
Vocal Fundamental Frequency	Age
Jitter	Height
Shimmer	Weight
Harmonic-to-Noise Ratio	UPDRS (Unified Parkinson's Disease Rating Scale)
Non-linear dynamical complexity measures	TUAG (Timed Up and Go test)
Signal fractal scaling component	Speed
Nonlinear measures of fundamental frequency variation	Gender

The author collected the data from two main sources for use in this investigation, with multiple features. Each of the datasets includes the feature indicating if the patient has Parkinson's or not, the sources are shown:

- *"UCI Machine Learning Repository (Human Speech Data) (Little, McSharry, Roberts, Costello, & Moroz, 2007)"*
- *"PhysioNet Physiologic Database (Demographic and Movement Data) (Frenkel-Toledo et al., 2005a) (Fren kel-Toledo et al., 2005b) (Hausdorff et al., 2007) (May berry et al., 2011)."*

Feature processing was performed on the datasets after each attribute from the demographic-movement-speech dataset was researched. The three steps followed for feature processing are: recognition setting, feature ranking and feature selection. In the recognition setting a process of unrecognizable characters and unnecessary data removal is taking place. Next, feature ranking evaluates the attributes to decide the optimal features in classifying whether the patient has the disease. This is done on both datasets by running evaluator algorithms and ranking which demographic, movement and speech factors are most important for the classification afterwards. To ensure consistent results, multiple attribute evaluator algorithms with implemented repetitions were used.

Data Analysis and Machine Learning

The objective of this investigation was a definite, binary classification problem distinguishing between two outcomes; positive to Parkinson's disease or negative. The machine learning process was made on data analytics software and raw programming IDEs using the Weka-Java API (Android device used for voice recordings). To determine if a patient has or not the disease classification type algorithms were used following two methods for the creation of the machine learning models; 400 repetitions onto 10-fold cross validation and a split of 70% training-set / 30% testing-set respectively. These models along with the selection process and feature ranking, ensured any over-fitting or mis-projected data were avoided.

For the analyzation of the machine learning models three main methods were conducted; data visualization (DV), model metric analysis (MMA) and external statistical testing (EST). MMA is used for identifying under-performing models by assessing their distinct model metrics. Four are the main metrics evaluating over fifteen machine learning models, from decision trees to random forests due to their ability to effectively assess their performance. The metrics are shown below:

- Percent accuracy (best closer to 100%)
- F-measure (best closer to 1)
- Logarithmic Loss (best closer to 0)
- Matthew's Correlation (best closer to 1)

Three main statistical significance tests were used for analysis, comparing the mean values for each model metrics, ranking them each respectively justifying which performed better than the rest. This is the observation to which models should basically be implemented for the application. The confidence level used for all the investigations was 95%.

Results

Demographic and Movement Data

As mentioned earlier, feature selection and selection ranking were the first performed task on the dataset. It confirmed which features were the most valuable during the classification process for Parkinson's disease. Dominating metrics viewed in the following table are the UPDRS, speed and TUAG showing their importance for determination of disease results. The UPDRS is a survey which reports the likelihood of individuals having Parkinson's (Fahn S, 2020).

Average Merit score determines the performance of the demographic-movement dataset attributes. The average rank also is representative of the attributes' performance, though to a lesser degree.

Table 3. Attribute's merit performance on demographic movement data (Ray, 2019)

Average Merit	Average Rank	Attribute
0.71 ± 0.009	1 ± 0	UPDRS
0.481 ± 0.018	2 ± 0	Gait Speed (m/s)
0.376 ± 0.018	3 ± 0	TUAG
0.143 ± 0.034	4 ± 0	Age (years)
0.077 ± 0.022	5.2 ± 0.4	Gender (binary)
0.040 ± 0.024	6.1 ± 0.7	Height (m)
0.022 ± 0.012	6.7 ± 0.46	Weight (kg)

The remaining features from the demographic movement data that ranked lower than the third placement were least effective for our needed classification, nevertheless they are essential for the creation of accurate machine learning models.

In order for the author to discover which of the fifteen models had inconsistent results and which were the best performing ones based on the metrics the model metric analysis (MMA) method was followed. First ranks the Locally Weighted Learning (LWL) at 98.8%, second Boosted Decision Stump at 98.69% while at third is the Decision Table at 98.62%. Table 4 represents how each model performed to model statistics respectively. The green boxes represent high-performing models compared to the red boxes representing relatively low-performing ones.

	DTable	Boostump	BoostTable	Htree	BayesNet	J48	Kstar	LMT	LWL	Neural Net	Naive Bayes	Random For	REP Tree	SGD	Stacking
Accuracy	98.62	98.68	98.2	92.25	85.55	84.63	81.05	98.25	98.8	82.93	96.13	88.94	84.09	86.35	56.03
Logarithmic Loss	0.12	0.06	0.06	0.12	0.33	0.27	0.37	0.09	0.06	0.33	0.15	0.27	0.27	0.35	0.5
Mathew's Correlation	0.97	0.97	0.97	0.93	0.71	0.7	0.63	0.97	0.98	0.66	0.93	0.78	0.7	0.73	0.63
F-Measure	0.99	0.99	0.98	0.94	0.88	0.87	0.83	0.98	0.99	0.85	0.97	0.9	0.87	0.88	0.72

Figure 2 Models performance for demographic movement data (Ray, 2019)

For the data visualization only the top three attributes from feature selection process were plotted since its purpose was the visual determination of individuals who have or do not have the Parkinson's disease. In the following 3D figure individuals diagnosed with the disease are marked with blue (#1 in scale) while those who are free of Parkinson's are in pink (#2 in scale).

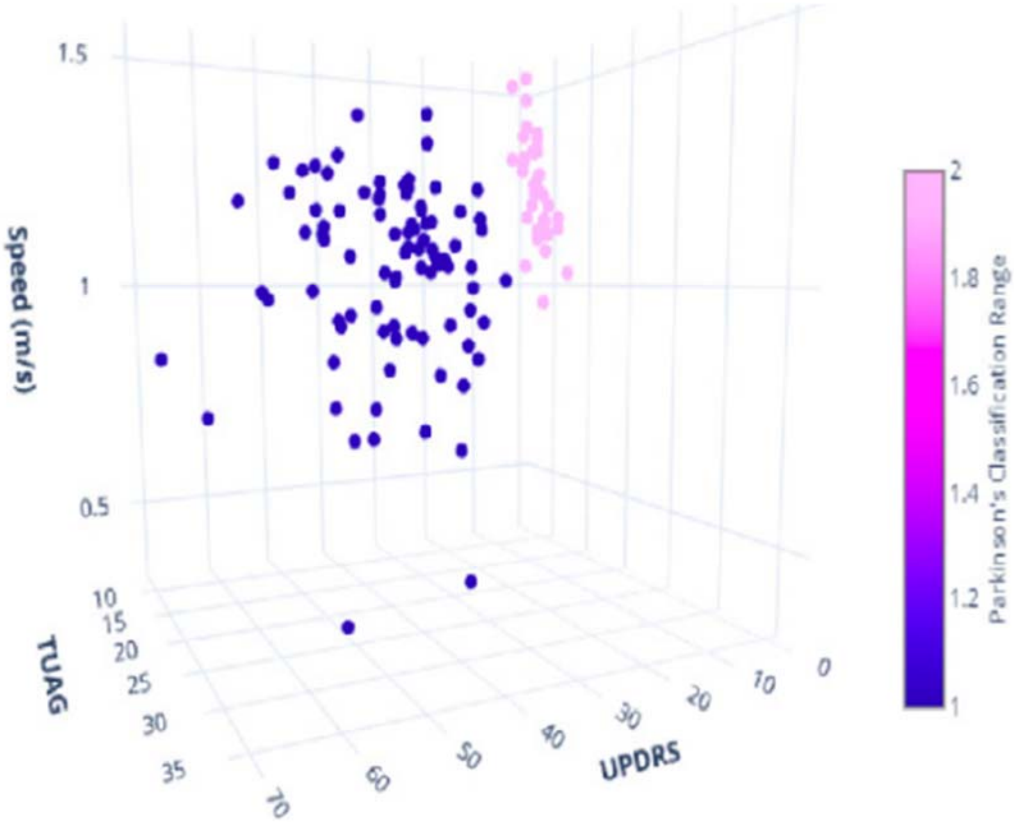


Figure 3. 3D visualization graph for demographic movement data

Human Speech Dataset

For the speech dataset, the same three stages including the model metric analysis, data visualization and external statistical testing were conducted through the machine learning procedure. From the feature ranking process, spread1, PPE and spread2 shown in the following table were discovered to be the most valuable attributes in determining the binary of diseased or not. Only three categories out of the nine different types of speech information were concluded as top features in conduction of the classification process. Non-linear measure of fundamental frequency variation, minimal fundamental frequency variation and shimmer from most to least important respectively. Once again, the remainder of attributes are necessary in order to reach the desired level of accuracy.

Table 4. Attributes merit performance on speech data (Ray, 2019)

Average Merit	Average Rank	Attribute
0.565 ± 0.012	1 ± 0	spread1
0.531 ± 0.013	2 ± 0	PPE
0.455 ± 0.016	3 ± 0	spread2
0.368 ± 0.010	5.6 ± 1.11	MDVP:Shimmer
0.380 ± 0.023	6.0 ± 1.90	MDVP:Flo(Hz)
0.384 ± 0.029	6.3 ± 2.72	MDVP:Fo(Hz)

0.366 ± 0.013	6.5 ± 1.50	MDVP:APQ
0.362 ± 0.014	7.3 ± 1.42	HNR
0.351 ± 0.010 9	9.5 ± 1.63	MDVP:APQ5

Similarly to the demographic and movement data, the model metric analysis was used across 15 distinct machine learning models for four different model metrics on the this speech dataset. Highest accuracy model is the "Stacking" at 94.6% with "KStar" and "Neural Network" models following at 91.8% and 91.3% respectively.

	DTable	BoostStump	BoostTable	Htree	BayesNet	J48	Kstar	LMT	LWL	Neural Net	Naive Bayes	Random For	REP Tree	SGD	Stacking
Accuracy	83.5897	85.1282	89.7436	75.3846	80	80.5128	91.7949	86.1538	84.1026	91.2821	69.2308	90.7692	86.1538	86.1538	94.5854
Logarithmic Loss	0.13	0.14	0.02	0.21	0.28	0.38	0.14	0.25	0.25	0.05	0.32	0.19	0.25	0.25	0.11
Mathew's Correlation	0.9	0.89	0.89	0.92	0.24	0.56	0.95	0.89	0.86	0.92	0.82	0.94	0.86	0.85	0.98
F-Measure	0.93	0.9	0.9	0.84	0.85	0.81	0.98	0.94	0.9	0.92	0.81	0.96	0.91	0.94	0.99

Figure 4. Models performance for speech data (Ray, 2019)

The author chose to visualize the speech in a very similar manner, plotting the variables of the top three mentioned categories determining once again which people do or do not have Parkinson's disease. The features met in the 3D diagram are the "HNR", λ variation and A variation. With a glance we can view less distinct results contrasting the demographic-movement dataset.

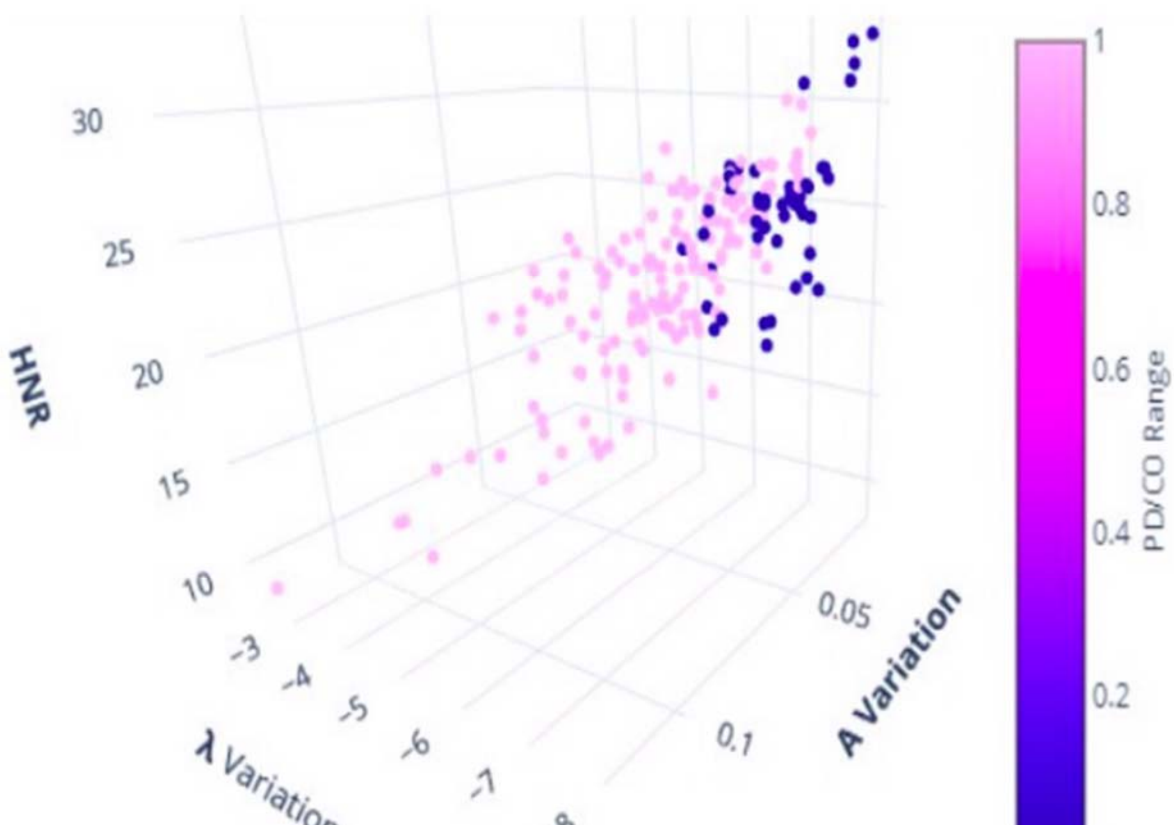


Figure 5. 3D visualization graph for speech data (Ray, 2019)

Discussion and performance

This investigation makes use of both statistical and machine learning approaches to determine and classify the neurological disease diagnosis on patients for Parkinson's disease. Two types of diagnostic models were created, one for demographic and movement data having an accuracy of 98.8% and one for the human speech data with its highest model reaching 94.6%. The author states that both these levels are highly more accurate than other pre-existing diagnostic procedures. This concludes to the fact of a expressively more accurate and cost-effective option a computational architecture offers compared to the 85% maximum accuracy level from the current neuropathologic standard responsiveness to medication. The standard procedures are conducted by the elimination of other diseases while new methods can be more reliable with a maximum mis-classification result just above 5%.

3.2 Similar Works

Another article worth mentioning, is Vladimir Despotovic's, Tomas Skovranek's and Christoph Schommer's research conducted in the University of Luxemburg in the computer science department in late 2019 (Despotovic, et al., 2020).

Similarly, their approach to the Parkinson's disease is based on speech impairments standing as one of the earliest symptoms. They approached their goal of identification of the diseased within their experimental study using Gaussian processes for regression and classification combined with Automatic Relevance Determination for their feature selection.

Two datasets were assessed in the study; the first detection dataset contains a field of biomedical voice measurements taken from 31 individuals where 23 are PD patients and 8 healthy, whereas the second dataset was conducted for telemonitoring, collecting biomedical voice measurements from 42 PD diagnosed patients for evaluation of their progress. The features used for determining the subjects' health are the ones mentioned at the topmost of this paper excluding the vocal fundamental frequencies for the telemonitoring of progress dataset.

Results

The performance of each models for the task was evaluated using sensitivity, accuracy and specificity as main measures. Through this experimental analysis of the two distinct tasks results obtained using the Gaussian processes are compared to various baselines, including decision tree ensembles (random forests, boosted and bagged decision trees) and Support Vector Machines for the identification.

The following figure shows the use of Gaussian Process classification with ARD (Automatic Relevance Determination) Matern 5/2 covariance function for the feature selection for the first dataset, showing that all but 5 features can be removed without affecting the overall performance.

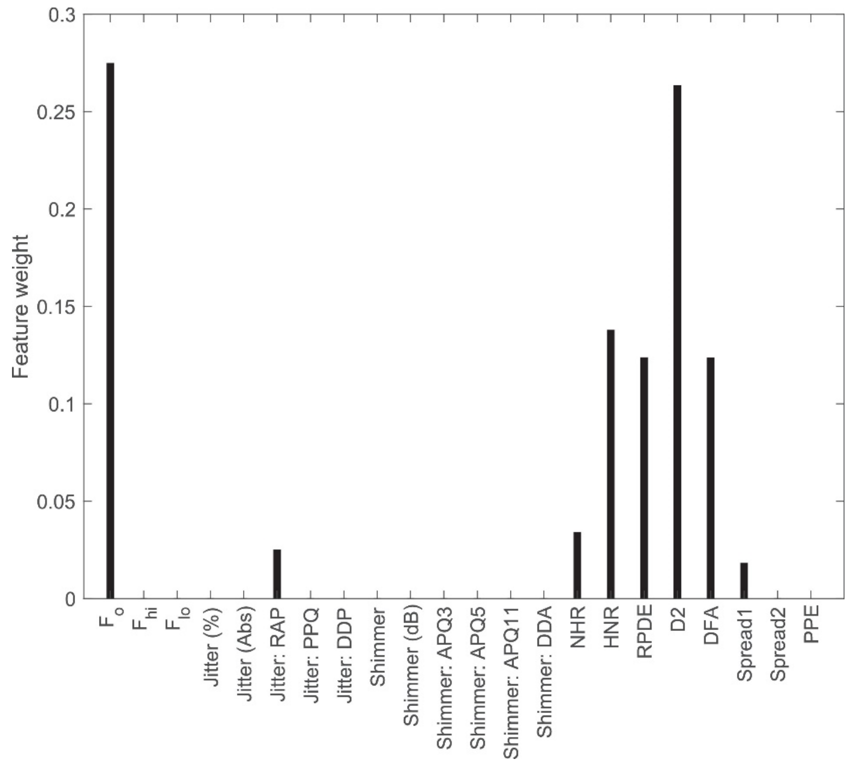


Figure 6. Classification with ARD, first speech dataset (Despotovic, et al., 2020)

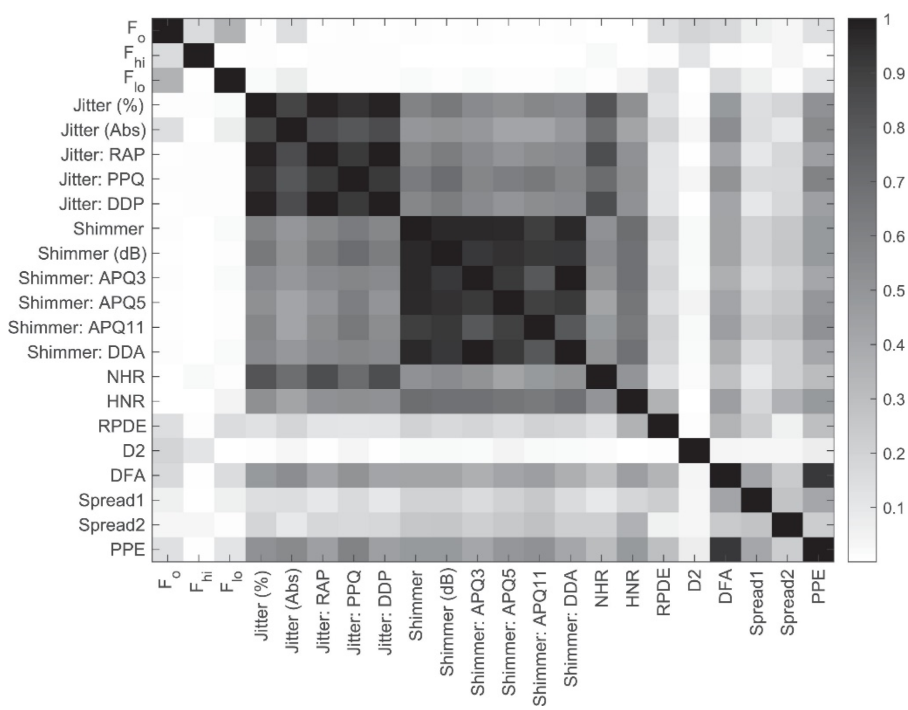


Figure 7. Classification with ARD, first speech dataset (Despotovic, et al., 2020)

Another figure follows to show the correlation between the voice features, where white color denotes no correlation whereas black color refers to perfect correlation.

Respectively, for the second dataset of telemonitoring the ARD is used for extracting the subset of relevant features shown in the figure below.

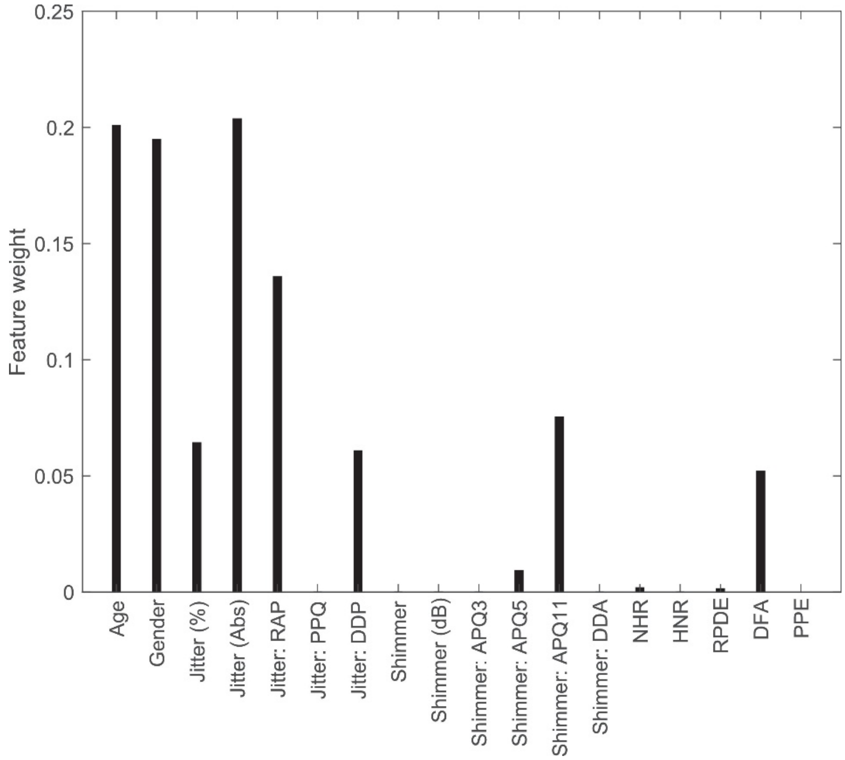


Figure 6. Classification with ARD, second speech dataset (Despotovic, et al., 2020)

Discussion and performance

The outcomes presented in the authors' paper show that Gaussian processes are a promising tool for this binary classification problem of PD identification also assessing the progress only from speech. Combined with the covariance functions with Automatic Relevance Determination, not only the performance of identification is improved but is also a great tool for the attribute selection. Since such a system's simplicity plays great role as well requiring only a patient's voice records, it can be implemented for monitoring and easily as a telemedicine application. With an accuracy of 96.92% of the Gaussian process, outperforming Support Vector Machines and decision tree ensembles it is concluded as the better methodology and approach for bringing new intuitions for early detection of symptoms for Parkinson's disease.

4. CONCLUTION AND PERSONAL VIEW

This research has been conducted in order to understand more about machine learning approaches and uses, giving the opportunity for better analyzing, comparing and manipulating multiple data at once. Moreover, it is a great approach for understanding possible difficulties or avoid similar to previous mistakes improving the overall process to be followed. Machine learning algorithms are great tools for making and finding new results but distinct methodologies and models are required for different kind of data at a time.

For this work, the human speech data is focused for the binary classification problem, identifying patients of having Parkinson's disease or not. By viewing these previous, while very recent, works, a better understanding of what the input features collected from individuals is reached to the expectation. Firstly, the approach through classification type algorithms is met giving results up to 94.6% accuracy leaving an estimate of 5% misclassification. Furthermore, Gaussian processes for regression and classification for determining an individual's health is used by other authors giving even greater results up to 96.92% accuracy.

Of course, these results are based on different datasets with possibly unbalanced data such as the ratio of diseased and non-diseased. Through these works, it is deduced that not all features are relevant or informative for our desired results, thus a feature selection model is best to be implemented before moving on to calculations.

Both results, are standing at a very high rate of accuracy compared to previously used methods standing around 85% accuracy for the declaration of the diseased. Personally, the Gaussian processes combined with Automatic Relevance Determination for their feature selection seem to be the ultimate method pre-processing data, since there are many different types of attributes in the dataset with most being differently scaled as variables. Comparatively, data analytics software and the use of classification algorithms used by cross validation and training/testing set splits seem very promising, accessible, easy to configure and experiment for the time given for implementation. In this case, the desired outcome is binary, which defines if the individual has or has not Parkinson's disease leading to a classification problem and so is the approach. It furthermore seems a great tool for using on human speech dataset, while compared to various baselines, including decision tree ensembles and Support Vector Machines. Finally, as of accuracy, the "stacking" machine learning model is also delivering great results while a 3D model would be optimal for visualizing the data. The dataset to be used is retrieved by UCI created by Max Little which includes voice samples from 31 people, where 23 of them are the diseased (Little, et al., 2008). Its attributes are shown in table 1.

Using speech data for the recognition of Parkinson's disease is a feature relatively unexplored. Although, current models show great accuracy of results that could certainly be used for an easier, faster and cost-effective method for helping patients' and doctors' progress in this field. It is definitely a matter worth working on, bringing new results that could eventually help many people's lives.

5. References

Anonymous, 2019. *WebMD.* [Online]
Available at: https://www.webmd.com/parkinsons-disease/guide/parkinsons-diagnosis-tests
[Accessed 2020].
Despotovic, V., Skovranek, T. & Schommer, C., 2020. *Speech Based Estimation of Parkinson's Disease Using Gaussian Processes and Automatic Relevance Determination,* https://www.sciencedirect.com/science/article/pii/S0925231220304318: ScienceDirect.
Education, I. C., 2020. *IBM.* [Online]
Available at: https://www.ibm.com/cloud/learn/machine-learning
[Accessed 12 2020].
Elkouzi, D. A., 2019. *Parkinson's Foundation.* [Online]
Available at: https://www.parkinson.org/understanding-parkinsons/what-is-parkinsons
[Accessed 12 2020].
Fahn S, E. R., 2020. *Theracycle.* [Online]
Available at: https://www.theracycle.com/resources/links-and-additional-resources/updrs-scale/
[Accessed 2020].
GEN Corporation, 2020. *GEN - Genetic Engineering and Biotechnology News.* [Online]
Available at: https://www.genengnews.com/news/advancing-parkinsons-disease-diagnosis-with-machine-learning/
[Accessed 12 2020].
Little, M. A., McSharry, P. E., Hunter, E. J. & Ramig, L. O., 2008. *UCI, Parkinsons Dataset, Suitability of dysphonia measurements for telemonitoring of Parkinson's disease.* [Online]
Available at: https://archive.ics.uci.edu/ml/datasets/Parkinsons
[Accessed 2020].
MathWorks Corporation, 2020. *MathWorks.* [Online]
Available at: https://www.mathworks.com/discovery/machine-learning.html
[Accessed 12 2020].
McDermott, A., 2019. *healthline.* [Online]
Available at: https://www.healthline.com/health/parkinsons/early-onset?c=767899079616
[Accessed 2020].
Physiopedia, 2020. *Physiopedia.* [Online]
Available at: https://www.physio-pedia.com/Parkinson%27s_Disease:_A_Case_Study
[Accessed 12 2020].
Ray, S., 2019. A PREDICTIVE DIAGNOSIS FOR PARKINSON'S DISEASE THROUGH MACHINE LEARNING. *THE CANADIAN SCIENCE FAIR JOURNAL.*
Samal, D., 2020. *Kraggle.* [Online]
Available at: https://www.kaggle.com/debasisdotcom/parkinson-disease-detection
[Accessed 2020].
Senturk, Z. K., 2020. *Early diagnosis of Parkinson's disease using machine learning algorithms,* https://www.sciencedirect.com/science/article/pii/S0306987719314148: ScienceDirect.
Triarhou, L., 2013. *Dopamine and Parkinson's Disease. In Madame curie bioscience database.,* s.l.: s.n.

UCI Corporation, 2009. *UCI Machine Learning Repository: Data Sets.* [Online]
Available at: https://archive.ics.uci.edu/ml/datasets.php
[Accessed 12 2020].

6. BIBLIOGRAPHY

Savara, G. (2020). *Parkinson's Disease Exploratory Data Analysis and Prediction.* [online] Data Insight. Available at: https://www.datainsightonline.com/post/parkinson-s-disease-exploratory-data-analysis-and-prediction.

Mathur, R., Pathak, V. and Bandil, D. (2018). Parkinson Disease Prediction Using Machine Learning Algorithm. *Advances in Intelligent Systems and Computing*, pp.357–363.

Gour, R. (2020). Using *Python to Detect Early Onset Parkinson's Disease*. [online] Medium. Available at: https://codeburst.io/using-python-to-detect-early-onset-parkinsons-disease-b89651b0ed3 [Accessed 21 Dec. 2020].

M. VENKATESWARA, R., T. Panduranga, V. and SVGK Dowluru, K. (2013). *Intelligent Parkinson Disease Prediction Using Machine Learning Algorithms*. [online] Research Gate. Available at: https://www.researchgate.net/publication/262817615_Intelligent_Parkinson_Disease_Prediction_Using_Machine_Learning_Algorithms

Almeida, J.S., Rebouças Filho, P.P., Carneiro, T., Wei, W., Damaševičius, R., Maskeliūnas, R. and de Albuquerque, V.H.C. (2019). Detecting Parkinson's disease with sustained phonation and speech signals using machine learning techniques. *Pattern Recognition Letters*, [online] 125, pp.55–62. Available at: https://www.sciencedirect.com/science/article/pii/S0167865519301163 [Accessed 21 Dec. 2020].

Karapinar Senturk, Z. (2020). Early diagnosis of Parkinson's disease using machine learning algorithms. *Medical Hypotheses*, 138, p.109603.

Nilashi, M., Ibrahim, O. and Ahani, A. (2016). Accuracy Improvement for Predicting Parkinson's Disease Progression. *Scientific Reports*, 6(1).