A Predictive Diagnosis for Parkinson's Disease Through Machine Learning

Shounak Ray

Age 16 | Calgary, AB

Currently, there is no fully reliable or accurate way to determine whether an individual has Parkinson's disease. Moreover, 90% of clinically-confirmed Parkinson's disease cases are idiopathic, suggesting the extent to which a more accurate and reliable diagnosis would be beneficial. This neurological disorder proffers a tremendous financial burden to the patient in terms of initial diagnostic and treatment costs. In this study, human demographic, movement, and speech data was analyzed to determine if an individual has Parkinson's disease, thus resulting in a binary classification problem. Two datasets encompassed these categories: one for demographics and movement and another for speech data. Machine learning and statistical testing were conducted on the two datasets individually. Over 30 different machine learning models, from lazy-based to tree-based, were analyzed through visualizations, model metric analysis, and external statistical testing. Upon in-depth exploration of the dataset and the multiple models, an Android application was created in order to prove the merits of each machine learning model. The application extracts users' demographic, movement, and speech data – both through manual input and artificial intelligence components, such as automatic speech recognition and model optimization. Ultimately, after rigorous statistical testing procedures, the locally weighted learning (LWL) was the best demographic-movement model (accuracy = 98.8%) and an ensemble model was the best human speech model (accuracy = 95.5%). Cumulatively, the machine learning framework founded on demographic, movement, and speech data suggests a more accurate, time-efficient, and cost-effective gold standard for Parkinson's disease diagnosis.

INTRODUCTION

There is an absence of time-efficient, cost-effective, and biologically accurate diagnoses for Parkinson's disease (PD) in the world, especially since doctors confound PD with other Parkinsonian syndromes (Goldman et al., 2018). Though some particular diseases have accurate diagnoses based on patients' biochemical responses, other recurrent diseases are difficult to detect early based on biochemical data and composition. For example, diagnoses for widespread ailments such as PD do not rely on firm biochemical data for determining whether the patient has the neurodegenerative disease. Instead, medical professionals depend solely on physical tests focusing on impaired movement and diminished hand-eye coordination for early detection. DatSCANs[1] and SPECT[2] analyses are also used to determine if a patient has PD. However, these are relatively cost-inefficient and inaccurate (Rizek, Kumar, & Jog, 2016). MRI[3], CT[4], and PET[5] scans are used not to directly diagnose PD. Rather, they are used to eliminate the possibilities of other confounding diseases (Rizek et al., 2016). Doctors do not actually determine whether a patient has Parkinson's; rather, they follow a process of elimination framework: by eliminating the likelihood for other diseases, they conjecture that the individual has PD. Additionally, results from these examinations have very low correlations with the presence of PD, as many other diseases also have the same symptoms. For instance, anemia and hypoglycemia also have symptoms of dizziness and impaired movement. Some unintended outcomes of such coping medications for these diseases include confusion, hallucinations, delusions, mood swings, and psychological changes (Saria & Zhan, 2018). An early detection system for Parkinson's will allow doctors to begin rehabilitative treatment early on, so the patient no longer has to consume medication for an extended period of time.

From a financial standpoint, it has been determined that widespread, inaccurately diagnosed diseases, such as PD, cost the

[1] Synonymous to most Ioflupane I123 Injections
[2] Single-photon Emission Computed Tomography
[3] Magnetic Resonance Imaging
[4] X-Ray Computed Tomography
[5] Positron Emission Tomography

US government as much as twenty-five billion dollars and affect more than ten million people worldwide (Marras et al., 2018). Moreover, PD is expected to affect nearly one million citizens in America by 2020 and 1.2 million citizens in America by 2030. Overall, the number of people with PD has already doubled with respect to its prevalence in 1978. Additionally, medication alone for PD costs $2,500 per year with therapeutic surgery (Marras et al., 2018).

Previous studies have utilized machine learning to develop an effective diagnosis for heart failure (Mallya, Overhage, Srivastava, Arai, & Erdman, 2019) and cancer (Wong & Yip, 2018), among other diseases. Machine learning technology has also been used to identify brain tumours by analyzing images of the human brain, (Sakai & Yamada, 2019) and to detect retinal disease in laboratory settings (De Fauw et al., 2018).

Some other organizations have also used machine learning for disease diagnosis; however, they have specifically cautioned that some inherent pitfalls in the machine learning algorithms may deliver inaccurate results (Park, 2018). For instance, researchers have suggested that machine learning techniques used to determine presence of disease have not always been successful in developing clinically validated diagnostic methods. In other words, some machine learning techniques such as classification and stochastic backpropagation are prone to over-fitting – not delivering a justified disease diagnosis.

On the other hand, other scientists in the machine learning community have suggested that machine learning techniques for diagnosis offer novel approaches to verify some unexplained phenomena in contemporary medicine – such as the relationship between demographics and disease prevalence (Sugai, Nomura, Gilmour, Stevens, & Shibuya, 2018). Thus, the introduction of a novel machine learning diagnosis for such diseases will not only reduce medical pecuniary losses but will also proffer a cost and time-efficient solution widely applicable in society.

Given the current constraints in Parkinson's disease diagnosis, the main question of this investigation is: how accurately and efficiently can an automatic machine learning model, which analyzes historical and live user demographic, movement, and speech data, diagnose Parkinson's disease as compared to existing, traditional procedures? The introduction of a novel machine learning diagnosis for Parkinson's disease will not only reduce medical pecuniary losses, but also proposes a cost and time efficient solution which is universally applicable.

METHODOLOGY
As previously mentioned, the project was divided into stages one, two, and three, each for distinct procedures in the investigation. The procedures of each of these stages are explicated below. For both datasets, the individuals were stratified across whether they had PD, early-stage or late-stage diagnosis, and geographical location. The geographical locations were equally balanced between North America, South America, India, and West Africa in order to sustain valid results.

Table 1. Complete List of Human Speech, Demographic, and Movement Features. The statements listed under the "Human Speech" column are simply names of the feature categories. Truly, there were many more sub-features under these headings which were used in the confirmatory procedures. Therefore, confirmation bias was diminished by ensuring the each of the features were distinct from each other.

Human Speech (n = 4485)	Demographic-Movement (n = 3700)
Vocal Fundamental Frequency	Age
Jitter	Height
Shimmer	Weight
Harmonic-to-Noise Ratio[6]	UPDRS[7]
Non-linear dynamical complexity measures	TUAG[8]
Signal fractal scaling component[9]	Speed
Nonlinear measures of fundamental frequency variation	Gender

Stage 1: Data Collection and Feature Processing
The data used in this investigation was accumulated from two main sources, as shown below, with multiple features, as shown in Table 1:

• 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) (Frenkel-Toledo et al., 2005b) (Hausdorff et al., 2007) (Mayberry et al., 2011).

In addition, each of the datasets includes a variable that indicates if a patient has Parkinson's or not, namely:

• PD[10], CO[11]

Expressed in set notation, this means that the FOI could be one of two values: PD represents that the patient has Parkinson's disease; whereas, CO represents "control," meaning the patient does not have PD.

After each of the features in the demographic-movement-speech (DMS) dataset were researched, feature processing was performed on both datasets. There were three steps in feature processing: recognition setting, feature ranking, and feature selection.

Recognition setting deleted unrecognizable characters from the DMS data – any character which was not a comma. These

---

[6]Abbreviated as HNR and further known as Harmonicity metric
[7]Abbreviation for Unified Parkinson's Disease Rating Scale
[8]Abbreviation for Timed Up and Go test
[9]Referred to as DFA value
[10]Parkinson's disease patient
[11]Control patient

were deleted from the data because the comma-separated value (CSV) and ARFF file formats were used for data analytics. So, a data file with non-comma characters would not be processed correctly by the ML program.

Next, feature ranking was performed on both datasets; different attribute evaluator algorithms were run on the DMS data files to determine the optimal features in classifying the FOI. In other words, the different algorithms ranked which demographic, movement, and speech factors were most important in determining if a patient is classified as PD or CO – the goal of the investigation. Though a single attribute evaluator algorithm would suffice for feature ranking, multiple were implemented to ensure consistent results (Blum & Langley, 2002). They are listed below as methods of the Weka-Java library:

weka.attributeSelection.ReliefFAttributeEval

.ClassiferAttributeEval

.ClassifierSubsetEval

.CorrelationAttributeEval

.PrincipleComponents

The weka.attributeSelection.PrincipleComponents attribute selection algorithm is also referred to as Principle Component Analysis (PCA). Feature selection was also performed as part of the feature processing stage, however this is further discussed in "Stage 2" as it is directly related to ML process.

Stage 2: Data Analysis and Machine Learning
For this investigation, the ML problem was a nominal, binary classification problem; the objective is to distinguish between two distinct outcomes under the FOI: PD or CO.

The main materials used in this investigation were a standard work computer, an Android device, and data from online repositories. The ML was conducted on data analytics software and raw programming IDEs using the Weka-Java API.

Classification algorithms were used to determine if the patient has PD and two methods were used to create the machine learning models: 10 fold cross validation (CV) across 400 repetitions and 70% training-set/30% testing-set split (TTS).

Both these methods were implemented to solidify the results of the investigation. The CV and TTS methods, along with the feature ranking and selection processes, ensured that over-fitting and mis-extrapolation of the DMS data were avoided (Ng, 1997). Further analysis of the model files, once they were created, also determined if they over-fitted, by investigating different performance metrics for each machine learning model.

There are three main methods through which the ML models were analyzed: model metric analysis (MMA), data visualization (DV), and external statistical testing (EST). The purpose of the MMA is to identify under-performing models by evaluating their different model metrics. Over fifteen distinct ML models, from

decision trees to random forests, were evaluated over 4 main metrics, as shown below. These four metrics were specifically due to their ability to effectively assess the performance of machine learning models, as opposed to other metrics (Model evaluation: quantifying the quality of predictions , n.d.).
   • Percent accuracy (closer to 100%, the better[12])
   • Logarithmic Loss (closer to 0, the better)
   • Matthew's Correlation (closer to 1, the better)
   • F-measure[13] (closer to 1, the better)

Percent accuracy was the metric analyzed to assess the effectiveness of the ML models because it is a direct measurement of how accurate the model is during the classification process. In addition, the loss, correlation, and F-measure are more specified metrics that allowed the models to be validated (Goutte & Gaussier, 2005). For example, a ML model with a high percent accuracy but equally high logarithmic loss means that the ML model is not reliable.

The DV process was conducted to further validate the ML models. Unlike the MMA and EST, the DV process does not directly relate to the ML models themselves. Instead, the original demographic, movement, and speech feature sets were visualized, in a 3D scatter-plot environment, in order to look for clustering of PD and CO instances. A relatively distinct cluster of PD and CO individuals would be in accordance with the "good" model statistics potentially received in the MMA. On the other hand, less distinct clusters would not justify the same, potentially "good" model statistics. This method, despite its advantages, is not enough to entirely justify the performance of each of the machine learning models since it is a rough overview. For additional validation, the EST procedures were implemented.

The EST is essentially a group of distribution analysis and statistical significance tests. This is necessary to eliminate the potential drawback of any statistical mishap in the MMA. For instance, high percent accuracy and logarithmic loss scores may simply be due to an error during statistical analysis; the significance tests allowed for us to differentiate between different ML models. The confidence level used for all the investigations was 95% ($\alpha = 0.05$).

Three main statistical significance tests (and test statistics) as presented below were used for analysis:
   • 2-sample Kolmogorov-Smirnov (KS) test
   • Student's t-test
   • Welch's t-test

Each of these significance tests compares the mean values for each of the machine learning model metrics, ranking each of the models. This ranking ultimately distinguishes which models justifiably performed better than others. This offers insight regarding which ML models should ultimately be implemented in the An-

---

[12]"Better" refers to the fact that the ML model is more reliable and accurate in binary classification
[13]Also known as "F1-score"

droid application. Moreover, three different tests are conducted to bolster the results achieved in the MMA and DV sub-stages. The fundamental architecture and paradigms governing the ML models were also investigated to have a comprehensive understanding of how the ML models classify a patient as having PD. This step allowed for the optimization of the ML models to generate the best performance.

Stage 3: Data Extraction and App Development

Once the ML models were created, the Android application was created. The Android application allows for a greater scope for user interaction, and thus greater DMS data collection. This DMS data is ultimately fed into incremental ML models. From the app development perspective, speech recognition APIs and Google Voice Analysis APIs were called upon to execute stage 3 of the project.

Additionally, the mobile application will deliver a live diagnosis for PD to the user. Android Studio was the platform where the app was created, which used the Weka-Java API for machine learning integration. The original training data is imported into the Android Application upon which the ML procedures are performed.

Three narrow[14] artificial intelligence (AI) methods are also present in the Android application:
- Automatic Speech Recognition (ASR)
- Speech Feature Extraction
- Automatic Model Picker

The ASR and extraction methods simply allow the app to recognize the user's voice and harvest human speech features (such as those shown in Figure 2). In contrast to the MMA in Stage 2 of the project, over thirty ML models were trained and the best one was picked, in terms of model metrics (such as those present in the MMA) and metric test-statistics (such as those conducted in the EST in Stage 2).

From a computational perspective, the CPU times required to run the models are also analyzed, through the Android Studio integrated development environment, in order to create the most efficient prediction system for the user.

RESULTS

Demographic and Movement Data

First, feature selection ranking and selection was performed on the dataset. This step demonstrated which variables were the most important during the PD classification process. UPDRS, speed, and TUAG were the top three most important metrics in determining whether an individual may have PD - as expressed in Table 2.

The UPDRS is a questionnaire which reports the likelihood of individuals having Parkinson's; this metric will be explored in

[14] Narrow AI refers to AI which can participate in just one specialized task

Table 2. Feature selection representation in the demographic-movement dataset. The performance of the demographic-movement dataset attributes were determined by the "average merit" score in the first column. The average rank also is representative of the attributes' performance, though to a lesser degree.

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)

more detail in the Discussion section. The speed simply refers to how fast the patient walks in meters per second. The TUAG test is a more quantitative measure of general human mobility: the patient starts sitting, walk to a point 3 meters away, and returns to his or her chair (Barry, Galvin, Keogh, Horgan, & Fahey, 2014). The time taken for this action is recorded as the TUAG score, which was the third most important metric in determining whether an individual has PD under the demographic-movement dataset.

Although the remainder of the features ranked below position three were much less effective in determining whether an individual has PD, these were nevertheless essential in creating accurate ML models. Holistically, the purpose of the feature selection process was simply to determine which factors were the most important in determining whether an individual has PD.

As shown in Table 3, fifteen different ML models were created via the MMA method and assessed according to the previously mentioned four model metrics. The MMA was used in order to discover which models had inconsistent results based on the metrics: no such models were identified. The best performing ML models in terms of accuracy were LWL (Locally Weighted Learning) at 98.8%, Boosted Decision Stump at 98.69%, and Decision Table at 98.62%.

The purpose of the DV was to determine a visual clustering of individuals who do and do not have PD. The top three features found in the feature selection process were plotted and distinct clusters of those with and without PD were found. This distinct clustering, as depicted in Figure 1, further justifies the high accuracy of the ML models in the Demographic-Movement dataset.

Once again, the EST allows us to differentiate between top performing models according to their model statistics. For instance, multiple sets of two models were compared based on metrics such as F-measure and percent accuracy to determine which model is better performing and is statistically significant

Table 3. Demographic-Movement Dataset: ML Model Metric Analysis. This represents how each of the ML models performed with respect to the model statistics. The green areas represent well-performing models while red areas represent relatively poor-performing models. The green and red highlights are determined by the distributions in each of the rows; in other words, for each of the model metrics.

|	DTable	Boostump	BoostTable	Htree	BayesNet	J48	Kstar	LMT	LWL	Neural Net	Naïve 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
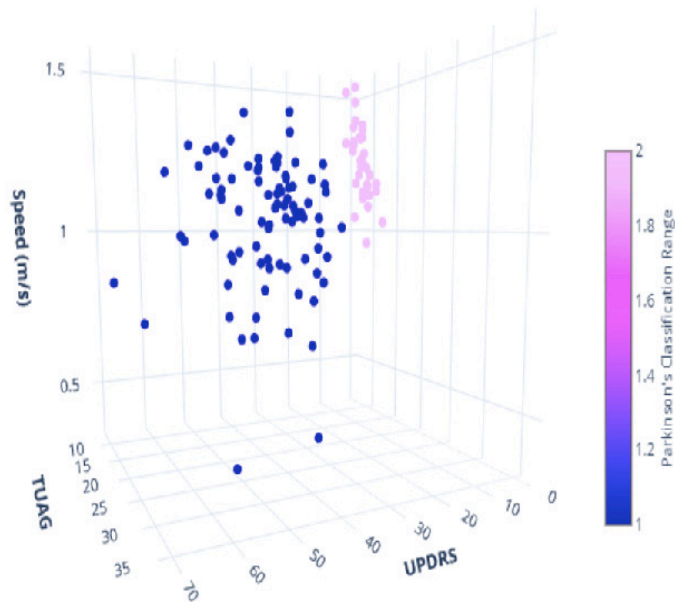
Figure 1. Demographic-Movement Top-Parameter Space. This was used to identify PD and CO distinctions.

as opposed to the comparison mode.

Furthermore, different histograms of the model metrics were created in order to visualize their distributions; a few distributions are depicted in Figure 2. For the demographic-movement dataset and the human speech dataset, the significance tests were performed to differentiate between the similar-looking distributions[15]. The reason there are distributions of values rather than single numbers for each metric-model cell is because the classification process was run many times: this led to a range of values.

For the following tests (Kolmogorov-Smirnov (Table 4), Student's-t (Table 5), and Welch's-t (Table 6)), P-values greater than $\alpha = 0.05$ do not represent a statistically significant difference and are presented more red. Similarly, P-values lower than $\alpha = 0.05$ indicate statistically significant results, and are presented in green. Additionally, for each of these tests, the green cells represent low KS P-values, which indicates higher chance of rejecting the null hypothesis. In this case, the null hypothesis is if the left-cell model is better than the top-cell model or if the top-cell model is better than the left-cell model. Conversely, the rela-

tively redder areas demonstrate higher P-values, failing to reject the null hypothesis. Ultimately, each of the models were ranked through these tests.

In conclusion, the locally weighted learning was the highest performing model in terms of the different model statistics: percent-accuracy, logarithmic loss, Matthew's correlation, and F-measure.

Human Speech Dataset

The same three stages were conducted on the human speech dataset: the MMA, DVs, and EST procedure.

In the feature ranking process, spread1, PPE, and spread2 were discovered to be the most important features in determining whether an individual has PD, as expressed in Table 7. These features are examples of some of the dynamical complexity measures analyzed in the human speech.

Among the nine different types of speech information collected, three of the categories were determined to have the most impact. In order from most important to least important, non-linear measure of fundamental frequency variation, shimmer, and minimal fundamental frequency variation were determined as the top features during the classification process.

Like the Demographic-Movement dataset, the remainder of the features which had a rank lower than three were less important in determining whether an individual has PD. Nevertheless, they were necessary to reach high levels of accuracy.

Once again, as shown below in Table 8, the MMA was conducted across 15 different ML models for 4 different model metrics on the human speech dataset.

According the MMA, the "Stacking," or ensemble, ML model had the highest accuracy at 94.6%. Consecutively, the "KStar" and the "Neural Network"[16] were the next best performing models with accuracy levels of 91.8% and 91.3% respectively.

Variables from the top three mentioned categories were plotted on the 3D scatterplot to determine a clustering between people who do and do not have PD. As shown in Figure 3, a cluster is demonstrated between the PD and CO groups. Though

---

[15] 60 histograms were created in total for further EST for each dataset, since 15 models were analyzed across 4 metrics.

[16] This model is a class of feed-forward artificial neural networks (fANNs).

Table 4. Heatmap of the 2-sample KS test results in demographic-movement dataset. In this scenario, F-measure was compared across the different models, via the 2-sample Kolmogorov-Smirnov test, to look for statistical significance at a 95% confidence level.

	LWL	Boosted Stump	Decision Table	Boosted Decision Table	Logistic Model Tree	Naïve Bayesian Model
LWL		0.07	0.06	0.06	0.07	0.04
Boosted Stump	0.07		0.04	0.04	0.02	0.01
Decision Table	0.06	0.04		0.08	0.06	0.06
Boosted Decision Table	0.06	0.04	0.08		0.05	0.03
Logistic Model Tree	0.07	0.02	0.06	0.05		0.06
Naïve Bayesian Model	0.04	0.01	0.06	0.03	0.06	

Table 5. Heatmap of the Student's t-test results in demographic-movement dataset. In this scenario, logarithmic loss was compared across the different models, via the Student's t-test, to look for statistical significance at a 95% confidence level.

	LWL	Boosted Stump	Decision Table	Boosted Decision Table	Logistic Model Tree	Naïve Bayesian Model
LWL		0.06	0.08	0.06	0.06	0.03
Boosted Stump	0.06		0.04	0.04	0.02	0.02
Decision Table	0.08	0.04		0.08	0.06	0.05
Boosted Decision Table	0.06	0.04	0.08		0.05	0.05
Logistic Model Tree	0.06	0.02	0.06	0.05		0.06
Naïve Bayesian Model	0.03	0.02	0.05	0.05	0.06	

Table 6. Heatmap of the Welch's t-test results in demographic-movement dataset. In this scenario, Matthew's correlation was compared across the different models, via the Welch's t-test, to look for statistical significance at a 95% confidence level.

	LWL	Boosted Stump	Decision Table	Boosted Decision Table	Logistic Model Tree	Naïve Bayesian Model
LWL		0.08	0.07	0.07	0.07	0.02
Boosted Stump	0.08		0.03	0.03	0.04	0.02
Decision Table	0.07	0.03		0.09	0.04	0.04
Boosted Decision Table	0.07	0.03	0.09		0.05	0.03
Logistic Model Tree	0.07	0.04	0.04	0.05		0.07
Naïve Bayesian Model	0.02	0.02	0.04	0.03	0.07	

it is less distinct than the corresponding visualization in the demographic-movement dataset, the visual separation is enough to justify the accuracy levels of the ML processes.

After the MMA and DV processes, the EST procedures were conducted in order to differentiate one top performing model from another. In other words, these procedures help rank the fifteen different ML models from "best" to "worst" as per their statistics. Once again, these statistical significance tests, the results of which are presented in Tables 9 - 11, followed the aforementioned guidelines and hypotheses at a 95% confidence level - as presented in Subsection Demographic and Movement Data. In conclusion, the ensemble/stacking ML model was the highest performing model in terms of the different model statistics: percent-accuracy, logarithmic loss, Matthew's correlation, and F-measure.

The only result analyzed in the app development process was the CPU time for running the app on an Android phone. This test was mainly conducted in order to determine the physical, computation efficiency of the application. The average build time for starting the application was about 0.231 milliseconds and the intermediate, artificial intelligence/machine learning processes took a maximum of 3000 milliseconds. The overall build time, however, was reduced by using single variant project sync, a method to speed up application processing (Hohmuth & Hermann, 2001).

DISCUSSION

Once again, the purpose of this investigation is to determine if an individual has PD based on the DMS data. Speech data is a feature which is relatively unexplored as an indicator of PD (Brabenec, Mekyska, Galaz, & Rektorova, 2017); however, it was critical in this investigation. This investigation utilized both pure statistical and machine learning approaches to classify an individual as PD or CO. Moreover, to phase the neurological disease diagnosis out to the public, an Android application was created which allows for a live diagnosis.

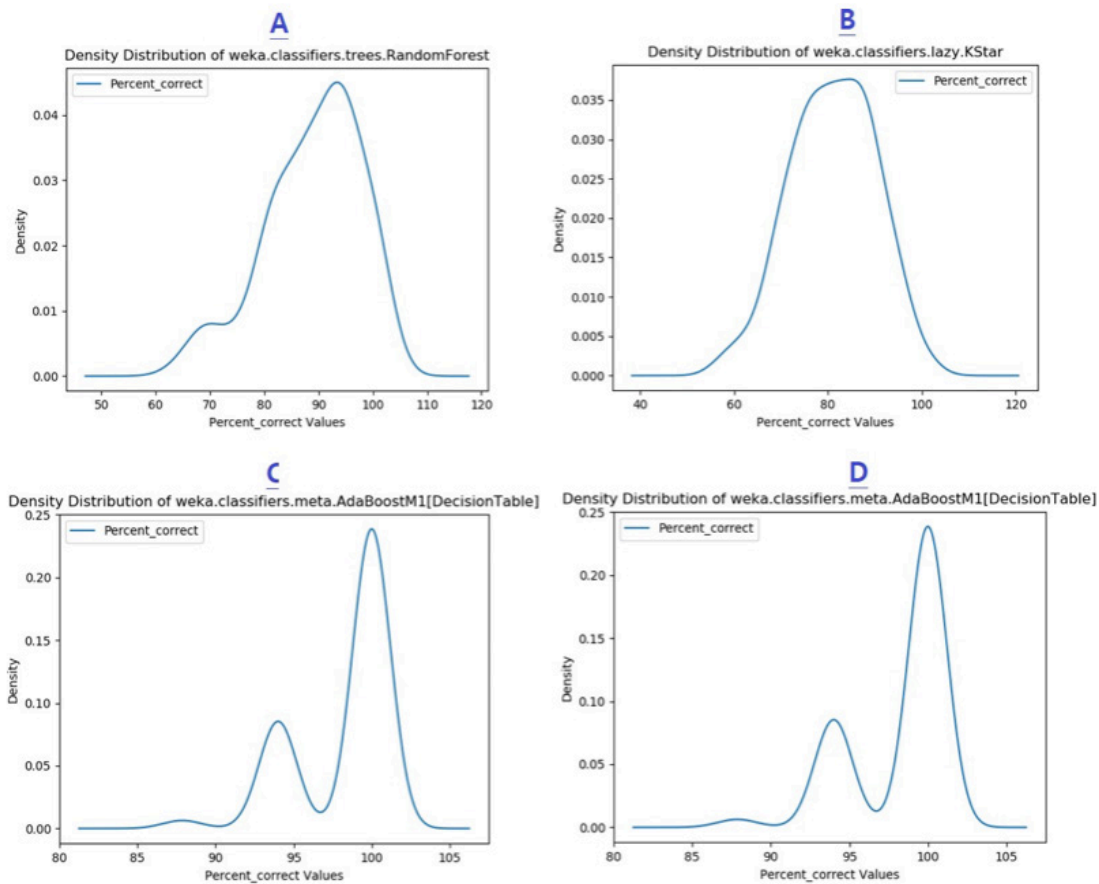Through comprehensive statistical validation and machine

Figure 2. The percent-correct metric distributions of four ML models in the demographic- movement dataset. The names of the models are shown in the individual header.   Panels A   and B show two similar metric distributions and panels C and D show additional, similar metric distributions.

Table 7. Feature selection representation in the human speech dataset. The performance of the human speech dataset attributes were determined by the "average merit" score in the first column. The average rank also is representative of the attributes' performance, though to a lesser degree. The MDVP and spread attributes simply represent other properties and ratios of the human voice which were analyzed in the classification process.

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	9.5 ± 1.63	MDVP:APQ5

learning, two types of diagnostic models were created: Category 1 and Category 2. Category 1 models refer to those created with demographic and movement data, of which the highest model had an accuracy of 98.8%. Category 2 models refer to those created with human speech data, of which the highest model had an accuracy of 95.5%. Both these accuracy levels are much higher than the existing diagnostic procedures. Consequently, the new models based off DMS data were proven to reduce the disease misdiagnosis rate from approximately 30% to 5%, offering a much more feasible and reliable diagnosis for PD.

The merits of the models and statistical analysis are dependent on how closely demographic, movement, and human speech factors correlate to the onset and presence of PD. Before conducting any statistical analysis regarding the DMS-based models, the features used in the study were analyzed to determine their correlations to PD. At its root, PD is caused by dopaminergic neurons in the substantia nigra pars compacta (Triarhou, 2013). This reduction in dopamine production ultimately affects

Table 8. Human Speech Dataset: ML Model Metric Analysis. This represents how each of the ML models performed with respect to the model statistics. The green areas represent well-performing models while red areas represent relatively poor-performing models.

	DTable	BoostStump	BoostTable	Htree	BayesNet	J48	Kstar	LMT	LWL	Neural Net	Naïve 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

Table 9. Heatmap of the 2-sample (KS) test results in human speech dataset. In this scenario, Matthew's correlation was compared across the different models to look for statistical significance at a 95% confidence level.

	LWL	Boosted Stump	Decision Table	Boosted Decision Table	Logistic Model Tree	Naïve Bayesian Model
LWL		0.07	0.08	0.06	0.07	0.04
Boosted Stump	0.07		0.04	0.05	0.02	0.01
Decision Table	0.08	0.04		0.08	0.02	0.06
Boosted Decision Table	0.06	0.05	0.08		0.05	0.03
Logistic Model Tree	0.07	0.02	0.02	0.05		0.06
Naïve Bayesian Model	0.04	0.01	0.06	0.03	0.06	

Table 10. Heatmap of the Student's-t test results in human speech dataset. In this scenario, F-measure was compared across the different models to look for statistical significance at a 95% confidence level.

	LWL	Boosted Stump	Decision Table	Boosted Decision Table	Logistic Model Tree	Naïve Bayesian Model
LWL		0.06	0.08	0.06	0.07	0.03
Boosted Stump	0.06		0.04	0.04	0.02	0.06
Decision Table	0.08	0.04		0.08	0.04	0.05
Boosted Decision Table	0.06	0.04	0.08		0.05	0.05
Logistic Model Tree	0.07	0.02	0.04	0.05		0.06
Naïve Bayesian Model	0.03	0.06	0.05	0.03	0.06	

Table 11: Heatmap of the Welch's-t test results in human speech dataset. In this scenario, logarithmic loss was compared across the different models to look for statistical significance at a 95% confidence level.

	LWL	Boosted Stump	Decision Table	Boosted Decision Table	Logistic Model Tree	Naïve Bayesian Model
LWL		0.08	0.05	0.07	0.05	0.02
Boosted Stump	0.08		0.03	0.03	0.06	0.02
Decision Table	0.05	0.03		0.08	0.04	0.04
Boosted Decision Table	0.07	0.03	0.08		0.05	0.04
Logistic Model Tree	0.05	0.06	0.04	0.05		0.07
Naïve Bayesian Model	0.02	0.02	0.04	0.04	0.07	

human cognition and, subsequently leads to physical and vocal impairment. Therefore, features in the DMS data were analyzed as potential predictors of PD and cross-validated with previous studies. For instance, many studies indicate that patients who are older (Reeve, Simcox, & Turnbull, 2014), shorter (Ragonese et al., 2007), and have a slower gait speed (Elbers, Van Wegen, Verhoef, & Kwakkel, 2013) have a significantly greater chance of having PD. Likewise, individuals have significant changes in their speech quality when they are beginning to show symptoms of the neurological disorder (Brabenec et al., 2017). Ultimately, all the features used in the machine learning process were confirmed to have some possible relationship to the presence of PD.

After this research process, the significantly higher accuracy levels found in this investigation – as compared to those currently present in the medical industry – were confirmed via rigorous statistical testing. These statistical techniques were used to validate the machine learning models. The feature ranking process determined that UPDRS was the top-ranked predictor of PD, although it had a merit score of 0.71 ± 0.009 at the 95% confidence level. This is important to note because, despite its

[17] Also referred to as sensitivity
[18] MDS-UPDRS in an abbreviation for the Movement Disorder Society Unified Parkinson's Disease Rating Scale, a metric intended to be more accurate than the UPDRS
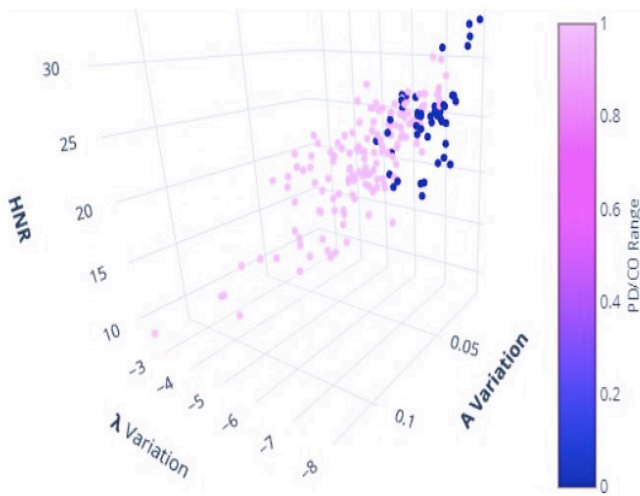
Figure 3. Human Speech Top-Parameter Space. This was used to identify PD and CO distinctions.

inaccuracy, UPDRS is the most commonly used diagnostic tool for PD (Kostek, Kaszuba, Zwan, Robowski, & Slawek, 2012). With accuracies17 no higher than 85% and 74.3%, as found in other studies, UPDRS and MDS-UPDRS18 respectively are individually unreliable metrics in determining whether an individual has PD (Starkstein & Merello, 2007) (Raciti et al., 2019). Despite the undependable performance of these metrics, UPDRS was still computationally determined to be a key feature in determining if a patient has PD in the demographic-movement dataset.

The 56% accuracy of ensemble method (stacking) in the demographic-movement dataset represents a rare case of how ensemble methods do not always produce higher prediction rates (Niculescu-Mizil & Caruana, 2012). It serves as a reminder than ensemble models should not simply be used due to their perceived efficiency as compared to standard, "regular" models. The lower-than-average accuracy level in the ensemble model was likely due to the superimposition of the noise in the datasets. For example, the demographic and movement dataset has features which
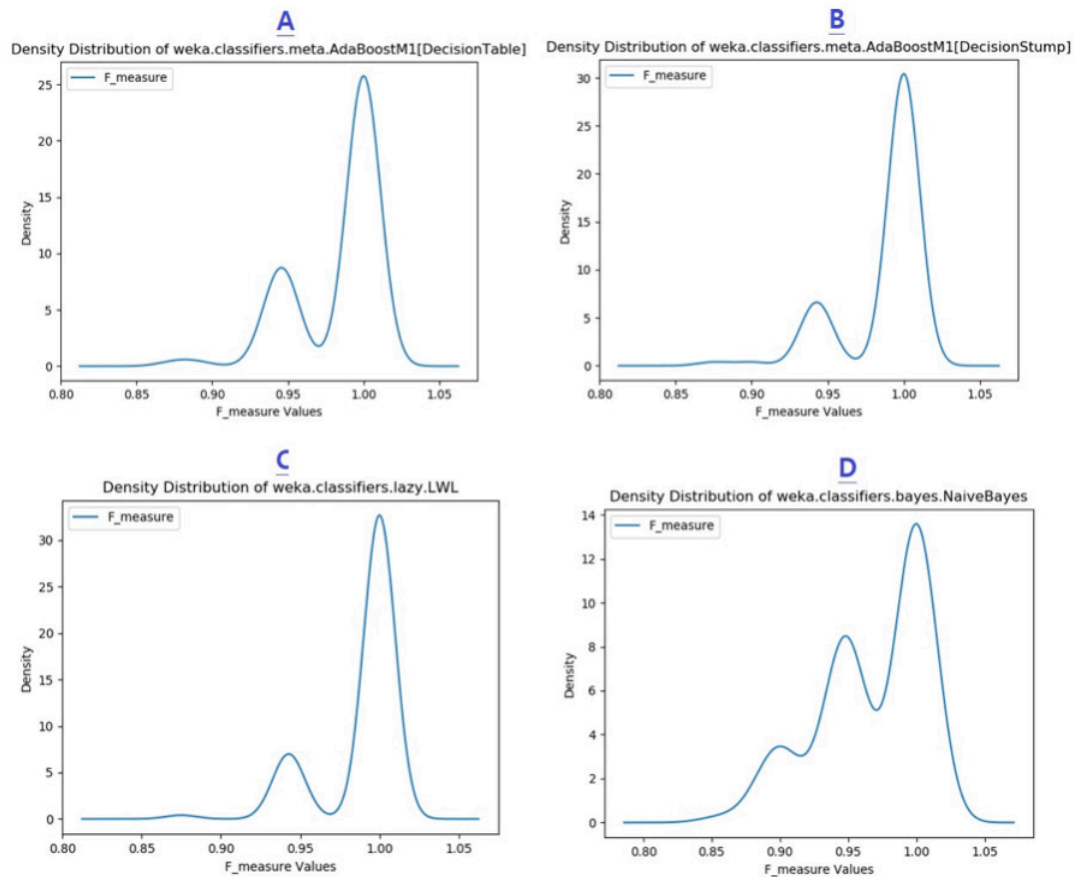

Figure 4. The F-measure metric distributions of four ML models in the demographic-movement dataset. The names of the models are shown in the individual header. Panels A and B show two similar metric distributions and panels C and D show additional, similar metric distributions.
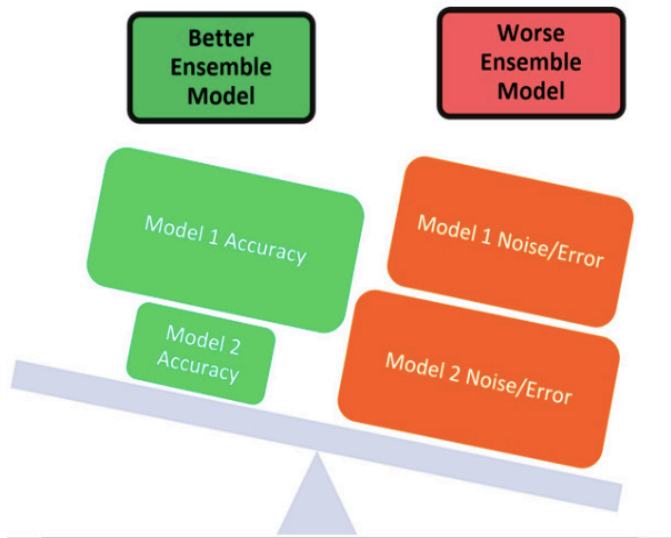
Figure 5. Ensemble Modeling Accuracy-Noise Trade-off in Category 1 Models. The balance represents how ensemble modeling had a magnified error when considering the demographic-movement dataset. The two models which were stacked had a cumulative noise-error value which outweighed the individual accuracies of the models. For example, though model 1 had a high accuracy, it was outweighed by model 2s higher noise-error impact.

are integral to the classification process, such as UPDRS, Speed, and TUAG. It also has other features, which were not very important in the classification process, such as age, gender, height, and weight. These simply increased the noise and error of the dataset. Now, when combining different models through the ensemble approach, both the accuracy and noise of the individual models are also aggregated in the finalized, stacked model (Orrell, 2005). In this circumstance, the noise and error of the models overpowered the corresponding accuracies, as depicted in Figure 5.

In contrast to the Category 1 stacked model, the Category 2 stacked model was the top performing model at 98.8% accuracy. By the same analogy, the high accuracies of the models had a greater impact than the combined noise-error value of the very same models. This resulted in an improved machine learning model for Parkinson's classification in the human speech dataset.

It is also important to note the role of bias and different feature-weightings in the machine learning methods. This condition is required to achieve justifiable results for the investigation. Firstly, unaltered and balanced data was used for the machine learning and statistical analysis procedures. Secondly, individual weightings in the feature set were not predetermined. For instance, the researcher did not explicitly program either Category 1 or Category 2 models to "focus more" on UPDRS or Speed when classifying whether a patient has PD. Every classification model used during training automatically decided which features were more

important than the others and are depicted in feature ranking results. Specifically, in this scenario, the relatively low bias in this dataset proves the verity of the statistical and machine learning results.

The MMA conducted on both datasets was used to identify inconsistent models. An example of an inconsistent model would be shown on the heatmaps as a mixture of red and green: this would represent that some of the model statistics are unexpected, rendering the model worthless. However, such inconsistencies were not identified in heatmaps (as shown in Table 3 and Table 8), meaning that all models were consistent and valid for further statistical analysis.

The DVs conducted for both datasets further confirmed the results of the machine learning models, specifically the accuracy levels. The distinct clusters shown in the 3D scatter-plots demonstrate that it is possible to roughly identify both PD and CO groups based on the top three metrics found in the feature ranking process. This is a significant step since it serves as a precursor for the machine learning classification process and allows us to achieve a greater understanding of the dataset. Without these visualizations, the machine learning process would be less intuitive and be treated more as a "black box," which would result in a superficial understanding of how the DMS data is being processed.

The last method of verification, the EST, allowed us to rank the different machine learning models and subsequently determine which models were "better" and why. As seen in the panel in Figures 2 and 4, the histograms of the model statistics – specifically percent accuracy and F-measure – visually seem quite similar: it is difficult to differentiate between the different graphs, and, therefore, the models. To distinguish between the different machine learning models, statistical significance tests were implemented ("Kolmogorov– Smirnov Test", 2008), especially the non-parametric Kolmogorov-Smirnov test. Here the reference distribution was one of the histograms and the comparison distribution was the other, similar-looking histogram in Figures 2 and 4. As previously stated, the purpose is to distinguish between the two histograms in order the find the better model. The P-values greater than $\alpha = 0.05$ indicated statistical significance, meaning that the comparison model was better than the reference model. Conversely, P-values lower than $\alpha = 0.05$ indicated insignificance, meaning that the comparison model was worse than the reference model. The comparative heatmaps for each of the three significance tests allowed us to create a relative matrix of P-values, providing a quick method to determine the top performing models in the DMS dataset.

Ultimately, this investigation offers a computational architecture which provides a significantly more accurate and cost-effective alternative to current procedures used to diagnose PD. Currently, there are four methods to determine if an individual has PD: physical tests, chemical tests, brain-imaging scans, and genetic testing. There is great discrepancy between the accuracy

levels of the UPDRS score, the most commonly used diagnosis for PD (Kostek et al., 2012), 2012). However, they do fall in range between

0.51 ≤ r² ≤ 0.71 (Brusse, Zimdars, Zalewski, & Steffen, 2005), much lower compared to the 95.5% and 98.8% accuracies achieved in this investigation. Moreover, 90% of clinically confirmed cases of PD are idiopathic (Ben-Shlomo & Sieradzan, 1995), meaning that the cause of the disease is unknown, and only 10% of cases are due to genetic causes (i.e. an extended family member was previously diagnosed with PD). Despite the high accuracy levels achieved in this study, many studies indicate that continuous testing is required to reduce the rate of mis-classification as well. As of now, a single diagnostic test – founded on chemical, physical, etc. exams – is the only diagnosis a person receives (Rajput, 1993). However, a Danish study indicated that the accuracy level of the diagnostic dramatically increases from 53% to 85%[19] if patients are re-diagnosed (Wermuth, Cui, Greene, Schernhammer, & Ritz, 2015). This is because repeated-testing reduces the scope of Type I[20] and Type II[21] error, consequently reducing the mis-classification rate. The reason why there is an absence of repetitive testing of PD is likely due to the high financial burden on the patient and health sector: it is simply unfeasible to conduct more than one neuropathologic test, such as the UPDRS test, for PD. Currently, the average cost of diagnosis is about $2,315[22] (Johnson et al., 2013) and treatment is about $1,000 per month (Brusse et al., 2005). Additionally, the burden of a Parkinson's patient, including diagnostic and treatment costs, amounts to a minimum of $30,000 (Muñoz, Kilinc, Nembhard, Tucker, & Huang, 2017). Melbourne's Howard Florey Institute recently developed a relatively cost-efficient diagnosis for PD, totalling to $4,000 (Low-cost Parkinson's disease diagnostic test a world first , 2007). However, this is only useful for the genetically-determined cases, not the larger proportion of idiopathic cases considered in this investigation.

Therefore, this machine learning framework will serve as a stronger prediction system for PD, however future re-diagnosis is recommended to ensure greater accuracy.

CONCLUSION

[19] This was the highest accuracy rate found for Parkinson's diagnosis as per current procedures

[20] The condition when the doctor says a patient has Parkinson's, but truly does not

[21] The condition when the doctor says a patient does not have Parkinson's, but truly does

[22] Refers to average price for consultation and any general combination of physical, chemical, imaging, and genetic testing

[23] Refers to any noticeable improvement in gait and speech after medication such as Carbidopa-Levodopa

The current neuropathologic gold standard for PD is responsiveness to medication[23] (Wermuth et al., 2015) with a maximum accuracy level of 85%. The procedures conducted in the gold standard do not truly diagnose a patient with Parkinson's disease: they only eliminate the possibility of other diseases (Rajput, 1993). The binary (PD or CO) machine learning framework developed in this investigation determines if a patient has Parkinson's disease using patients' demographic, movement, and speech data – rather than following an unreliable process-of-elimination framework. This investigation suggests that instead of conducting the multitude of rudimentary physical tests and relatively inaccurate chemical and brain-imaging procedures, an analysis of demographic, movement, and speech data – as per the DMS dataset features analyzed – is confirmed to be more accurate and reliable than the currently accepted gold standard (minimum accuracy = 95%, maximum mis-classification = 5%). To further reduce this mis-classification rate, repetitive testing is now a feasible option as there are no costs related to inputting DMS data into the free Android application.

REFERENCES

Barry, E., Galvin, R., Keogh, C., Horgan, F., & Fahey, T. (2014). Is the Timed Up and Go test a useful predictor of risk of falls in community dwelling older adults: A systematic review and meta-analysis. *BMC Geriatrics, 14* (1). doi: 10.1186/1471- 2318-14-14

Ben-Shlomo, Y., & Sieradzan, K. (1995). Idiopathic parkinsons disease: Epidemiology, diagnosis and management. *British Journal of General Practice, 45* (394), 261–268.

Blum, A. L., & Langley, P. (2002). Selection of relevant features and examples in machine learning. *Artificial Intelligence, 97* (1-2), 245–271. doi: 10.1016/s0004-3702(97)00063-5

Brabenec, L., Mekyska, J., Galaz, Z., & Rektorova, I. (2017). Speech disorders in Parkin- son's disease: early diagnostics and effects of medication and brain stimulation. Journal of Neural Transmission, 124 (3), 303–334. doi: 10.1007/s00702-017-1676-0

Brusse, K. J., Zimdars, S., Zalewski, K. R., & Steffen, T. M. (2005). Testing functional performance in people with Parkinson disease. Physical Therapy , 85 (2), 134–41. doi: 10.1093/ptj/85.2.134

De Fauw, J., Ledsam, J. R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S.,. . . Ronneberger, O. (2018). Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine, 24* (9), 1342–1350. doi: 10.1038/s41591- 018-0107-6

Elbers, R. G., Van Wegen, E. E., Verhoef, J., & Kwakkel, G. (2013). Is gait speed a valid measure to predict community ambulation in patients with Parkinson's disease? *Journal of Rehabilitation Medicine, 45* (4), 370–375. doi: 10.2340/16501977-1123

Frenkel-Toledo, S., Giladi, N., Peretz, C., Herman, T., Gruendlinger, L., & Hausdorff, J. M. (2005a). Effect of gait speed on gait rhythmicity in Parkinson's disease: Variability of stride time and swing time respond differently. *Journal of NeuroEngineering and Rehabilitation, 2* (23). doi: 10.1186/1743-0003-2-23

Frenkel-Toledo, S., Giladi, N., Peretz, C., Herman, T., Gruendlinger, L., & Hausdorff, J. M. (2005b). Treadmill walking as an external pacemaker to improve

gait rhythm and stability in Parkinson's disease. *Movement Disorders, 20* (9), 1109–1114. doi: 10.1002/mds.20507

Goldman, J. G., Holden, S. K., Litvan, I., McKeith, I., Stebbins, G. T., & Taylor, J. P. (2018). Evolution of diagnostic criteria and assessments for Parkinson's disease mild cognitive impairment. *Movement Disorders, 33* (4), 503–510. doi: 10.1002/mds.27323

Goutte, C., & Gaussier, E. (2005). A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation. In D. E. Losada & J. M. Fernández- Luna (Eds.), Advances in information retrieval (pp. 345–359). Berlin, Heidelberg: Springer Berlin Heidelberg. doi: 10.1007/978-3-540-31865-1 25

Hausdorff, J. M., Lowenthal, J., Herman, T., Gruendlinger, L., Peretz, C., & Giladi, N. (2007). Rhythmic auditory stimulation modulates gait variability in Parkinson's disease. *European Journal of Neuroscience, 26* (8), 2369–2375. doi: 10.1111/j.1460- 9568.2007.05810.x

Hohmuth, M., & Hermann, H. (2001). 2001 USENIX Annual Technical Conference Pragmatic nonblocking synchronization for real-time systems. In Usenix annual technical conference. Dresden, Germany.

Johnson, S. J., Kaltenboeck, A., Diener, M., Birnbaum, H. G., Grubb, E., Castelli-Haley, J., & Siderowf, A. D. (2013). Costs of parkinson's disease in a privately insured population. *PharmacoEconomics, 31* (9), 799–806. doi: 10.1007/s40273-013-0075-0

Kolmogorov–Smirnov Test. (2008). In The concise encyclopedia of statistics (pp. 283–287). New York, NY: Springer New York. doi: 10.1007/978-0-387-32833-1 214

Kostek, B., Kaszuba, K., Zwan, P., Robowski, P., & Slawek, J. (2012). Automatic assessment of the motor state of the Parkinson's disease patient: A case study. *Diagnostic Pathology, 7* (1). doi: 10.1186/1746-1596-7-18

Little, M. A., McSharry, P. E., Roberts, S. J., Costello, D. A., & Moroz, I. M. (2007). Exploiting nonlinear recurrence and fractal scaling properties for voice disorder de- tection. *BioMedical Engineering Online, 6* (23). doi: 10.1186/1475-925X-6-23.

Low-cost parkinson's disease diagnostic test a world first. (2007). EurekAlert! AAAS. Retrieved from https://www.eurekalert.org/pub releases/2007-02/ra-lpd022207.php

Mallya, S., Overhage, M., Srivastava, N., Arai, T., & Erdman, C. (2019). Effectiveness of LSTMs in Predicting Congestive Heart Failure Onset. arXiv e-prints, arXiv:1902.02443.

Marras, C., Beck, J. C., Bower, J. H., Roberts, E., Ritz, B., Ross, G. W., . . . Tanner, C. (2018). Prevalence of Parkinson's disease across North America. *npj Parkinson's Disease, 4* (1). doi: 10.1038/s41531-018-0058-0

Mayberry, K., Mancini, M., Manca, M., Ferraresi, G., Sensi, M., Cavallo, M., & Chiari, L. (2011). The effect of deep brain stimulation on gait asymmetry in parkinson's disease. *Gait & Posture, 33* (1), 1–2. doi: 10.1016/j.gaitpost.2010.10.006

Model evaluation: quantifying the quality of predictions. (n.d.). Scikit Learn. Retrieved from https://scikit-learn.org/stable/modules/modelevaluation.html

Muñoz, D. A., Kilinc, M. S., Nembhard, H. B., Tucker, C., & Huang, X. (2017). Evaluating the cost-effectiveness of an early detection of Parkinson's disease through innovative technology. *Engineering Economist, 62* (2), 180–196. doi: 10.1080/0013791X.2017.1294718

Ng, A. Y. (1997). Preventing Overfitting of Cross-Validation Data. In Icml '97 proceedings of the fourteenth international conference on machine learning (pp. 245–253).

Niculescu-Mizil, A., & Caruana, R. A. (2012). Obtaining Calibrated Probabilities from Boosting. arXiv e-prints, arXiv:1207.1403.

Orrell, D. (2005). Ensemble Forecasting in a System with Model Error. *Journal of the Atmospheric Sciences, 62* (5), 1652–1659. doi: 10.1175/jas3406.1

Park, S. H. (2018). Regulatory Approval versus Clinical Validation of Artificial Intelligence Diagnostic Tools. *Radiology, 288* (3), 910–911. doi: 10.1148/radiol.2018181310

Raciti, L., Nicoletti, A., Mostile, G., Bonomo, R., Dibilio, V., Donzuso, G., . . . Zappia, M. (2019). Accuracy of MDS-UPDRS section IV for detecting motor fluctuations in Parkinson's disease. *Neurological Sciences, 40* (6). doi: 10.1007/s10072-019-03745-2

Ragonese, P., D'Amelio, M., Callari, G., Aiello, F., Morgante, L., & Savettieri, G. (2007). Height as a potential indicator of early life events predicting parkinson's disease: A case-control study. Movement Disorders, 22 (15), 2263–2267. doi: 10.1002/mds.21728

Rajput, D. R. (1993).Accuracy of clinical diagnosis of idiopathic parkinson's disease. *Journal of Neurology, Neurosurgery & Psychiatry, 56* (8), 938–

939. doi: 10.1136/jnnp.56.8.938

Reeve, A., Simcox, E., & Turnbull, D. (2014). Ageing and Parkinson's disease: Why is advancing age the biggest risk factor? Ageing Research Reviews, 14 (1), 19–30. doi: 10.1016/j.arr.2014.01.004

Rizek, P., Kumar, N., & Jog, M. S. (2016). An update on the diagnosis and treatment of Parkinson disease. *Cmaj, 188* (16), 1157–1165. doi: 10.1503/cmaj.151179

Sakai, K., & Yamada, K. (2019). Machine learning studies on major brain diseases: 5-year trends of 2014–2018. *Japanese Journal of Radiology, 37* (1), 34–72. doi: 10.1007/s11604-018-0794-4

Saria, S., & Zhan, A. (2018, Jul). Us20180206775a1 - measuring medication response using wearables for parkinson's disease. Google. Retrieved from https://patents.google.com/patent/US20180206775A1/en

Starkstein, S. E., & Merello, M. (2007). The unified Parkinson's disease rating scale: Validation study of the mentation, behavior, and mood section. *Movement Disorders, 22* (15), 2156–2161. doi: 10.1002/mds.21521

Sugai, M. K., Nomura, S., Gilmour, S., Stevens, G. A., & Shibuya, K. (2018). Demographic and clinical factors associated with having ischemic heart disease as a multiple contributing causes of death among diabetes mellitus deaths in the united states and brazil. *Endocrine Abstracts, 56*, 384. doi: 10.1530/endoabs.56.p384

Triarhou, L. (2013). Dopamine and Parkinson's Disease. In Madame curie bioscience database.

Wermuth, L., Cui, X., Greene, N., Schernhammer, E., & Ritz, B. (2015). Medical Record Review to Differentiate between Idiopathic Parkinson's Disease and Parkinsonism: A Danish Record Linkage Study with 10 Years of Follow-Up. *Parkinson's Disease*, 2015, 1–9. doi: 10.1155/2015/781479

Wong, D., & Yip, S. (2018). Machine learning classifies cancer. *Nature, 555* (7697), 446–447. doi: 10.1038/d41586-018-02881-7

SHOUNAK RAY

Shounak Ray is an avid machine learning enthusiast who loves to discover its practical applications in the medical industry. As a keen programmer since grade 8, he has expertise in web development and Object-Oriented Programming. He has also used his technical knowledge to develop novel systems that alleviate stress for Alzheimer's patients and detect Parkinson's Disease with high, verified accuracy. With the world moving towards a data-driven future, Shounak Ray witnesses Big Data's potential to increase operational precision and enhance our understanding across multiple industries - especially the medical sector. Continuing his education, Shounak Ray would not only like to implement this Parkinson's diagnostic procedure in clinics, but also advocate data literacy in the public setting.