# Design and Validation of a Classification Model for Parkinson's Disease

Rahman Ajibade, Micky Nnamdi, Joyita Roy, Erin Shappell, and Michelle Warren
ECE 6254 - Spring 2022

## 1. Summary

Parkinson's Disease is one of the most common movement disorders, largely affecting older populations, with progressively developing symptoms caused by a slow breakdown of cells in the nervous system. Symptoms include hand tremors, gait impairment, and slurred speech to name a few. In this report, we will describe a method for training a machine learning model to predict the likelihood of a subject being a Parkinson's Disease (PD) patient based on a subject's gait and demographic data. A machine learned approach to detecting Parkinson's could assist doctors with early detection and possible mitigation of disease progression. Three different approaches were attempted: 1) Random Forest (RF) regression to achieve feature selection for training, coupled with Support Vector Classification with the reduced feature-selected dataset, 2) Convolutional Neural Network (CNN) with the entire dataset, and 3) Logistic Regression with the entire dataset. The dataset used to accomplish this is the open-source Physionet gait dataset, which includes 3 test instantiations of PD and control patients. Each patient has 8 sensors placed on the bottom of each foot, where force data may be obtained during a 2 minute walking period, hence the gait data. This data also includes demographic data of each subject, such as age, gender, walking speed, and the assessed Timed Up And Go score. The feature selection method turned out to leave out useful information for assessing PD likelihood, and therefore did not produce great results. The CNN provided a high precision of 99% for assessing high likelihood of PD in known PD patients, but a low precision of 49% for assessing that the likelihood for a control subject was low and thus performed poorly on identifying that the person does not have PD. The final assessment was that the Logistic Regression approach provided the highest accuracy when performed on test data at 87% accurate. Accuracy here is determined as correctly predicting a PD patient based on ground truth of diagnosis and/or correctly predicting that a person from the control group does not have PD.

## 2. Introduction

### *A. Project Background*

Parkinson's disease (PD) is a condition that is progressive in nature, with five main stages that leads to a breakdown of cells in the central nervous system [1]. Currently, the cause of Parkinson's is unknown, though it is speculated that it is most likely due to genetic and environmental factors; however, it should be noted that there is no concrete evidence of the specific instigators for disease onset [2]. A distinguishing feature of PD is the tremor, which usually occurs in a finger or foot that is generally at rest. This tremor generally occurs at about 4 to 6 cycles per second in a cyclic pattern. Tremors in the arm can occur at a higher frequency than the legs,(3-8 Hz as compared to 1-3Hz) [3]. Current PD treatments include brain stimulation surgery, which sends electric pulses through the brain through electrodes. It is hypothesized that akinesia (or bradykinesia) is the main contributor to the impairment of

movement in Parkinson's disease patients [4]. The current methods of detection of Parkinson's Disease are limited and not streamlined, leaving diagnosis up to the doctor's assessment of the patient's family history and performance in neurological and physical tests. Early detection of PD can help mitigate the progression of the disease to worse conditions where such drastic treatments are needed.

### *B. Existing Work*

The existing work in this realm mainly makes use of the Parkinson's Progression Markers Initiative (PPMI) dataset. The PPMI dataset is composed of observational clinical and longitudinal data of person's with a preexisting PD diagnosis, those of high risk, and those that are healthy. Research is in work to better understand how to use machine learning techniques to provide diagnosis for PD. Anila M's review on PD diagnosis using machine learning techniques outlines many different types of PD data, such as voice samples, speech data, and the HandPD dataset, which comprises handwritten exams from PD patients; and how they are used as inputs to many different machine learning models, ranging from simple algorithms such as SVMs and regression to more complex algorithms such as artificial neural networks, probabilistic neural networks (PNN), and deep belief networks [5].

### *C. Dataset*

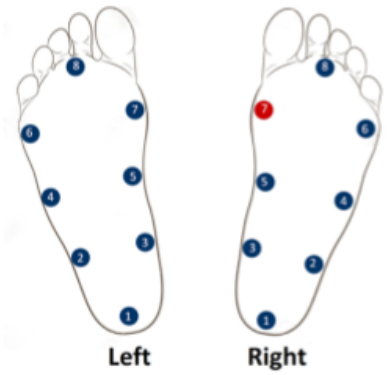


**Figure 1.** 8 sensors under the feet of participants (DhilipSanjay, Feb 2022)

The dataset used for analysis was gait data in PD patients from Physionet. The dataset includes 3 patients with Parkinson's disease and a control group of 73 healthy subjects (with 296 samples). The subjects walked for 2 minutes with 8 sensors under their feet, as can be seen in Fig 1, and the force-recordings were sampled at 100 Hz. The recordings also include two signals that reflect the sum of the 8 sensor outputs for each foot. Furthermore, it includes demographic information and measures of disease severity (i.e., using the Hoehn & Yahr staging and/or the Unified Parkinson's Disease Rating Scale).

### *D. Proposed Solution*

Our proposed solution is to design and validate a method by which we can accurately label a patient as having PD or not based on features of their movement. These features include, but are not limited to, walking speed and Timed Up and Go (TUAG) score: a measure of patient mobility. We chose to train and evaluate the performance of three models: the support vector classifier (SVC), convolutional neural network (CNN), and logistic regressor. We trained the SVM using a reduced version of the dataset and the CNN and logistic regressor with the full dataset.

### *E. Task Breakdown*

**Table 1.** Breakdown of team members that lead key tasks

| Task | Lead Team Member(s) |
|---|---|
| Background reading material | All team members |
| Data acquisition | Michelle |
| Feature Selection | Erin |
| Choosing Models | Micky, Erin, Joyita |
| Data Normalization (for logistic regression) | Rahman |
| Training the Models | Micky, Erin, Joyita |
| Testing/Validating the Models | Micky, Erin, Joyita |
| EDA (data visualization) | Michelle, Rahman |

## 3. Technical Approach

### *A. Feature Selection*

For our project, we chose to use Random Forest (RF) regression's built-in ranking of feature importances to aid in our selection of which features to use for our risk assessment model. During model fittings, the RF algorithm will assign an important value to each feature. This importance metric is calculated based on each feature's contribution to the pureness of the leaves of the decision tree. The importance of all the features adds to 1, so this metric may be used to rank features on the order of their importance. To determine which features best predict the severity of Parkinson's disease, we chose to train three RF regressors on each of the three provided PD scores: UPDRS, UPDRSM, and Hoehn & Yahr. Then, we compared the importance rankings of each model and selected the features that were consistently ranked highly as our reduced feature set. We found that all three models ranked the TUAG score and walking speed as the top two most important features (with average importance scores of 0.232 and 0.086, respectively), so we selected those two features to use for our initial model training. Table 2 summarizes the results of the individual RF regressors.

**Table 2**. Results from using random forest regression for feature importance ranking

| Model | Importance | |
|---|---|---|
| | TUAG | Walking speed |
| UPDRS | 0.201 | 0.069 |
| UPDRSM | 0.205 | 0.085 |
| Hoehn & Yahr | 0.289 | 0.104 |

### *B. Support Vector Classification with Reduced Dataset*

Based on the success of previous groups, we chose to train a support vector classifier (SVC) on the reduced dataset and used the provided patient labels (PD vs. control) as the ground truth labels. To optimize the classifier we trained three different models, each with a different kernel type: linear, radial basis function (RBF), and polynomial. A grid search was performed on each model to optimize the parameters C and $\gamma$ (and the degree for the polynomial kernel). A 70:30 train-test split was used.

The training set accuracy of the SVCs ranged from 65 - 71%, while the testing set accuracy ranged from 74 - 84%. The SVC using the RBF kernel performed the best; however, an 84% accuracy was lower than we desired. Because we were dissatisfied with these results, we aimed to (1) use all features in our dataset rather than a subset of features and (2) explore other models such as convolutional neural networks (CNNs) and logistic regressors. The results of each SVM are summarized in Table 3 with their respective confusion matrices in Figure 2.

**Table 3**. Results from an SVM using linear, RBF, and polynomial kernels.

| Parameter/Indicator | Linear | RBF | Polynomial |
|---|---|---|---|
| Accuracy on Training Set | 0.70 | 0.71 | 0.65 |
| Accuracy on Test Set | 0.84 | 0.84 | 0.74 |
| C | 10 | 10 | 100 |
| $\gamma$ | $1e^{-5}$ | 0.01 | 0.1 |
| Degree | N/A | N/A | 3 |

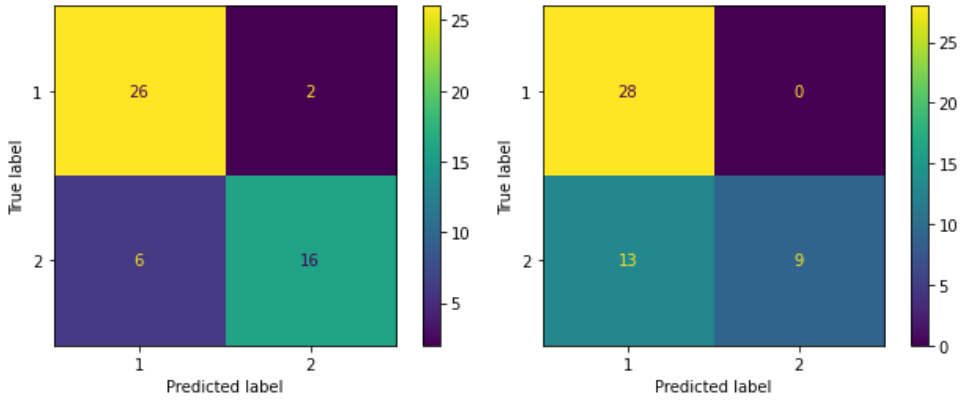


**Figure 2**. Confusion matrices summarizing the results of applying a trained SVM with linear (left), RBF (left), and polynomial (right) kernels to the test set. *Note that the confusion matrix for the linear and RBF kernels is the same.*

### *C. Convolutional Neural Network with Full Dataset*

We used a 1D convolutional neural network (1D-Convnet) to build a Neural Network classifier. The neural network consists of 1 convolution layer and 1 max-pooling layer, 1 dropout layer, flatten layer and 2 dense layers fully connected. The training was done using 100 epochs and SoftMax as the activation function.

We adopted the implementation of early stopping to monitor the model's performance during training and to prevent overfitting. If there is a constant decline in the performance of the validation set after 5 cycles, it will terminate the cycle. From Figure 3, we see the model does a decent job in training the data and there was no decline in its performance hence early stopping is not activated.

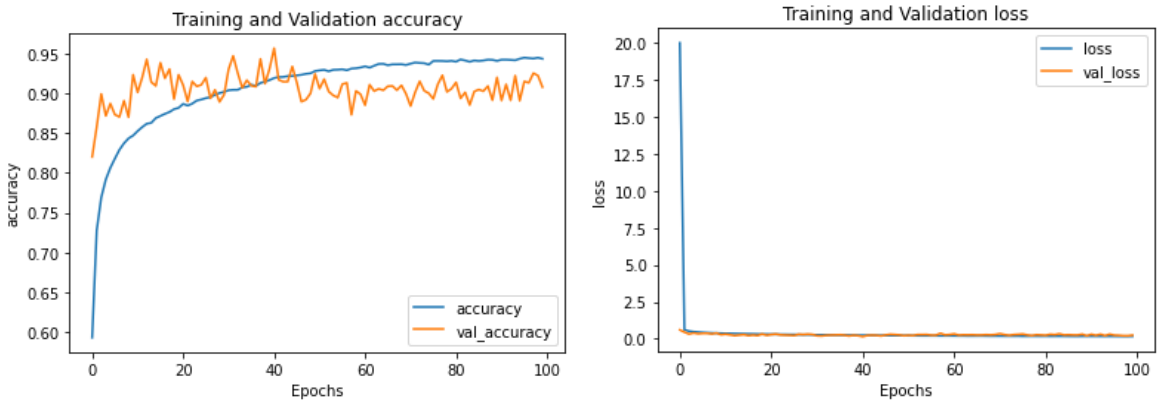


**Figure 3.** Graphs summarizing the training and validation accuracy (left) and loss (right) of the CNN.

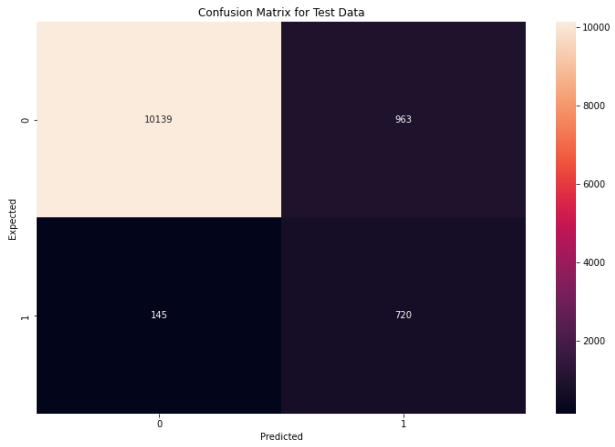


**Figure 4.** Confusion matrix summarizing the results of applying a trained CNN to the testing set.

Though the model does a decent job when you consider its accuracy, it does poorly predicting patients without Parkinson's disease when you consider evaluating its performance using precision. It achieves a precision of 43% on patients without Parkinson's disease and 99% on patients with Parkinson's disease. This performance can also be seen in the confusion matrix. Likely causes could be imbalanced data. Balancing this data before training could solve this problem.

### *D. Logistic Regression with Full Dataset*

The first step involved transforming the full dataset using statistical techniques: mean, standard deviation, skewness, kurtosis, median, maximum, and minimum. These are calculated for all 19 features and are used to compress the dataset. Subsequently, we scaled the values to be between 0 and 1 so that there are no inconsistencies between different units of measurement.

Prior to logistic regression, we used scikit-learn's train_test_split() function and followed the same split as prior models did. We chose to use logistic regression after transforming the dataset since it is one of the simpler models to implement, easy to interpret, and quick to train. We also used grid search to get an optimal value for C. Table 4 displays our results from logistic regression and performance is also measured using a confusion matrix (Figure 5).

**Table 4.** Logistic Regression Results

| Parameter/Indicator | Value |
|---|---|
| Accuracy on Training Set | 83.2 |
| Accuracy on Test Set | 87.0 |
| C | 5 |

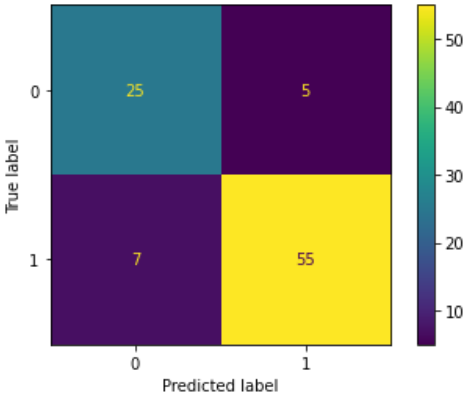


**Figure 5.** Confusion matrix summarizing the results of applying a trained logistic regressor to the testing set.

## 4. Conclusion and Future Work

In conclusion, we have trained a model to accurately identify a PD patient that may be used in a program for fast labeling and early detection of PD. Our Logistic Regression model performed the best at labeling PD patients and control subjects. While we initially used a Random Forest model to reduce the dataset before model training, we chose to use the full dataset to increase labeling accuracy in future model training and testing. While our CNN model also achieved high performance on labeling PD patients, it did not perform well when labeling control subjects; this suggests that the data must be balanced prior to training and a need to adjust for overfitting. Ultimately, the Logistic Regression model provides a clean and accurate method for detection and achieves the ultimate goals of the project.

## 5. References


[1] De Laum, L., Breteler, M. M. (2006). Epidemiology of Parkinson's disease. The Lancet Neurology, 525-535.

[2] Kalia, L. V., Lang, A. E. (2015). Parkinson's disease. Lancet, 896-912.

[3] Jankovic, J. (2008). Parkinson's disease: clinical features and diagnosis. Journal of neurology, neurosurgery and psychiatry, 368-376.

[4] Dauer, W., Przedborski, S. (2003). Parkinson's Disease: Mechanisms and Models. Neuron, 889-909.

[5] Anila M., Pradeepini, G. 2020. A Review on Parkinson's Disease Diagnosis using Machine Learning Techniques, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH and TECHNOLOGY (IJERT) Volume 09, Issue 06 (June 2020).

[6] DhilipSanjay, 2022. Human-Biomechanic-Analysis. https://github.com/DhilipSanjay/Human-Biomechanic-Analysis.git


## 6. Appendix

**Table A1**. Index of code used for all project components.

| Project Component(s) | Link to Colab Notebook |
|---|---|
| RF and SVM | https://colab.research.google.com/drive/1TYkZ3lpf0sq9rbjUEQ_2-B6o1UbfJ_uL?usp=sharing |
| CNN | https://colab.research.google.com/drive/1zoMI5ivvs-w_xiepMHEHVoB7B4m3WR9r |
| Logistic Regression | https://colab.research.google.com/drive/11dmTJZ3mYbvi2fjJires_0KH8h9_YXrE?usp=sharing |