Classification of subjects with Parkinson's Disease using Gait Data Analysis

Yash Mittra
Department of Information Technology
Guru Gobind Singh Indraprastha University
New Delhi, India
mittrayash@gmail.com

Vipul Rustagi
Department of Information Technology
Guru Gobind Singh Indraprastha University
New Delhi, India
vipulrustagi11@gmail.com

***Abstract*— Gait Analysis of Parkinson's Disease (PD) Patients and Control Objects has been analyzed to show differences in PD Patients and Control Objects. Using data provided by Phisonet's Gaitpdb database (in which 8 sensors have been applied to each foot of the subjects to calculate the Vertical Ground Reaction Forces (VGRF)), data compression has been performed using 7 statistical functions to get a representative image of the data. The statistical functions namely Minimum, Maximum, Mean, Median, Standard Deviation, Skewness, and Kurtosis have been used to compress over 3 million tuples into 310 tuples. Finally, various Machine Learning techniques have been applied to the transformed dataset to perform detection of Parkinson's Disease. The classification has been performed using Logistic Regression, Decision Trees, Random Forest, SVM (Linear Kernel), SVM (RBF Kernel), SVM (Poly Kernel) and k-Nearest Neighbours.**

Experiments with Principal Component Analysis for data compression have also been performed and their incompetence (with reasons) has been stated.

***Index Terms*—Parkinson's Disease (PD), Control Objects (CO), Gait Analysis**

I. Introduction

Parkinson's Disease is a degenerative disorder which targets the central nervous system. It affects the dopamine-producing neurons found in the brain which hampers movement, primarily, of limbs in the body. There is no standard test to diagnose Parkinson's Disease, a condition that affects up to one million people in the US [1]. Symptoms develop slowly over the years which include tremors in hands, unbalancing while walking and even an altered taste in smell in a few cases as per Parkinson.org. As the disease advances the symptoms typically become more severe and weakening. The disease also causes non-motor symptoms which often appear before a person experiences motor symptom and can prove to be more troublesome for some.

Non-motor symptoms include fatigue, excessive saliva, constipation, vision and dental problems and lack of facial expressions. Another interesting observation in PD patients is their inability to generate high force levels in limbs during locomotion. Recent studies and experiments have shown that people suffering from this disease are substantially slow in initiating a force production and are unable to produce smooth forces as per Pietro Mazzoni et al [2]. The motor behavior Laboratory of The University of Wisconsin-Madison examined the preparation and production of forces in Parkinson's Disease and found that patients, both young and elderly, could only generate force levels that were a percentage of their maximum.

Parkinson's Disease affects the brain in an intensely adverse manner. Within the brain, the major pathological change is progressive degeneration of neurons in the pars compacta of the substantia nigra, one of the nuclei that constitute the basal ganglia (BG). These neurons normally transmit dopamine to another BG nucleus, the striatum, but their degeneration leads to dysfunction of these neuronal circuits that include the BG and motor cortical areas [2].

According to Parkinson's Association of Carolinas, approximately 60,000 Americans are diagnosed with Parkinson's disease every year and an estimated 7-10 million people worldwide are currently living with this disease. The Incidence of Parkinson's disease increase with age, but an estimated four percent of people with PD are diagnosed before the age of 50.

This project aims to identify patients suffering from Parkinson's Disease by analyzing gait data of PD patients. The data used has been obtained from Physionet Gait Analysis Database. The database consists of data regarding 93 patients with idiopathic PD and 73 healthy control objects. It consists of VGRF of subjects as they walk for approximately 2 minutes. Every subject has a total of 16 sensors, 8 on the bottom of each foot (as depicted in the figure), which calculate the force in Newton as a function of time.

This data is then transformed through the application of statistical functions. This significantly reduced data has then been scaled between (-1,1) to increase data consistency and reduce computational power required while also keeping singularity of the data intact. The new reduced data has then been used to produce models through machine learning algorithms.

II. Methodology

This section deals with different methodologies carried out in this project such as Data interpretation, Data Transformation, Application of Classification Techniques along with other miscellaneous tasks (Time Series Analysis, Most Prominent Features). Fig 1 depicts the bifurcation of the two disparate pathways taken in the process of research.

Prior attempts to use Principal Component Analysis (PCA) to reduce the features, and hence, in turn, reduce the computation power required rendered unsuccessful. A justification for the failure for PCA is that the number of tuples was still over 4 million. Hence, the computational power required to apply models to our data was still all the same.

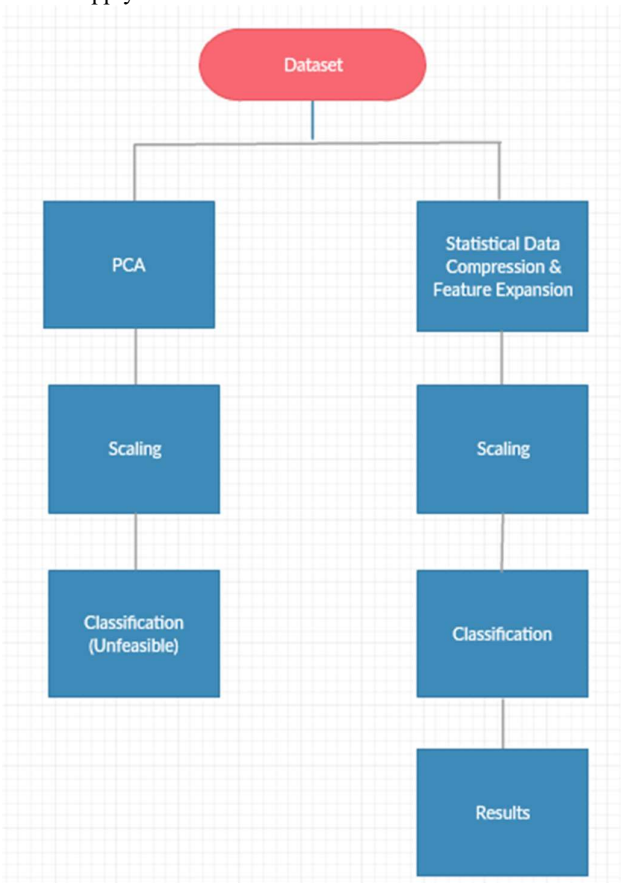

Fig 1: Experimented Methodologies in a Flow Chart

By using scikit-learn's [3] implementation of PCA, we were able to create a Random Forest classifier with an accuracy of 75%. This creation of this model took about 4 hours on Intel Core i5-5200U CPU @ 2.20GHz (4 CPUs) with 8 Gigabytes of RAM. However, this classifier was made using 1/5th of the CO data and 1/10th of the PT data. Most models could not be applied on the whole dataset because of the still high computational power required and the time complexity of the algorithms like SVM (RBF Kernel) being high.

III. Data Interpretation

Data has been collected by 3 studies [4]: Galit Yogev et al (Ga), Hausdorff et al (Ju), Silvi Frenkel Toledo (Ju), one of whose initials are present in each datafile. The second part of the data file name consists of either 'Co' or 'Pt' which represents Control Object and Parkinson Disease patient respectively. There are a total of 310 files each with 12,118 tuples of data. As a whole, they amalgamate into over 3 million tuples with each tuple having 19 features. 16 of these features are sensor values, one is Time and the rest two are the total force values exerted by the left and right foot. Fig 2 defines the segmentation of the columns across the dataset. Fig 3 defines the names of the columns in each of the segments.

```
Each line contains 19 columns:

Column      1:   Time (in seconds)
Columns   2-9:   Vertical ground reaction force (VGRF, in Newton) on each of 8
                 sensors located under the left foot
Columns 10-17:   VGRF on each of the 8 sensors located under the right foot
Column     18:   Total force under the left foot
Column     19:   Total force under the right foot.
```

Fig 2: Description of Features in The Dataset

Time	Left Foot	Right Foot	Total Force
Time(sec)	VGRF_left_s1	VGRF_right_s1	Total_force_left
	VGRF_left_s2	VGRF_right_s2	Total_force_right
	VGRF_left_s3	VGRF_right_s3	
	VGRF_left_s4	VGRF_right_s4	
	VGRF_left_s5	VGRF_right_s5	
	VGRF_left_s6	VGRF_right_s6	
	VGRF_left_s7	VGRF_right_s7	
	VGRF_left_s8	VGRF_right_s8	

Fig 3: Classified Version of the Feature Set

IV. Time Series Analysis

Toledo et al. [5] state that the ability to maintain a steady gait rhythm is impaired in patients with Parkinson's Disease. A visual of the Time Series Analysis has been created to highlight the differences in the VGRF force patterns in PD patients and Control Objects. Fig 4A highlights the differences in Left Foot Stride Force Patterns in Control Objects and PD while Fig 4B does the same for the Right Foot Stride Force Patterns for subjects with dissimilar weights. Fig 5A and Fig 5B highlights these differences in Left and Right Foot Stride Force Patterns respectively for subjects with similar weights. The visuals have been created using Matplotlib [6]. Through this time series, it is clear that there exist significant differences in the walking patterns of a PD patient and a Control Object. Gait abnormalities in PD include shortened stride length [7], [8], a dyscontrol of stride frequency [9], and postural instability.

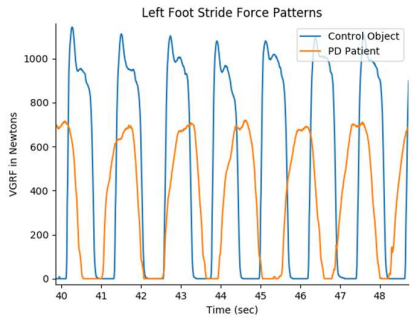

Fig 4A: Time Series Analysis of Subjects with Dissimilar Weights (Left Foot Stride Force Patterns)

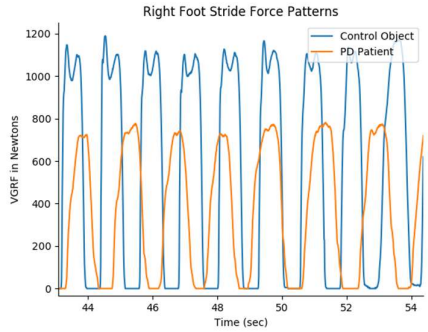

Fig 4B: Time Series Analysis of Subjects with Dissimilar Weights (Right Foot Stride Force Patterns)

Here, the difference in the force is mainly due to the CO chosen for the time series analysis weighed 83 kg, whereas the PD patient chosen weighed 50 kg. The differences in weights have been deliberately chosen to prevent clustering in the time series. However, the heel strike pattern differences will continue to be evident even when we choose to use subjects with similar weights.
This can be visualized as under.

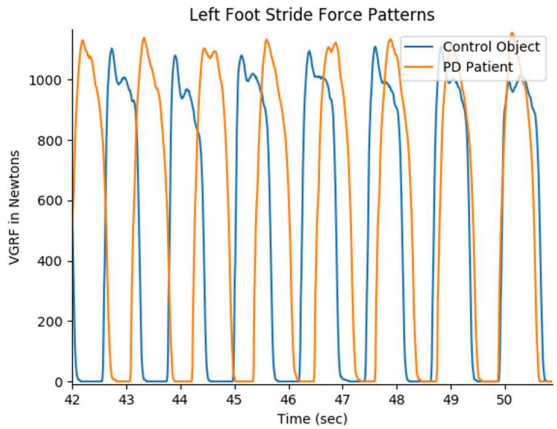

Fig 5A: Time Series Analysis of Subjects with Similar Weights (Left Foot Stride Force Patterns)

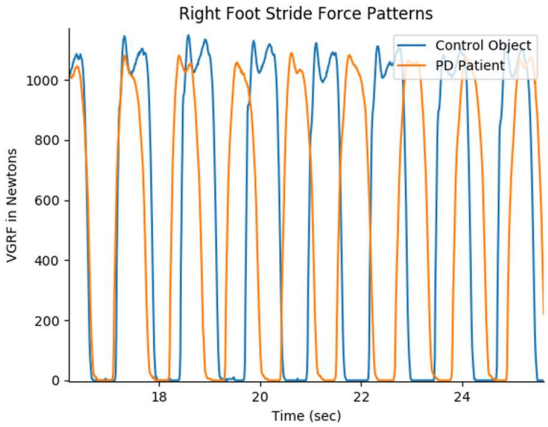

Fig 5B: Time Series Analysis of Subjects with Similar Weights (Right Foot Stride Force Patterns)

Time Series Analysis of Subjects with Similar Weights
In this case, the Control Object weighed 83 kg and the PD Patient weighed 82 kg. The VGRF heel strike patterns continue to depict the difference.

V. Data Transformation

Such a plethora of data cannot be used directly to train models. The computational power required is enormous. Times like these call for the need of feature transformation. This is a technique which can bring together data in an optimal format. In this project, we look at a statistical approach to transform our data with techniques - Minimum, Maximum, Mean, Median, Standard Deviation, Skewness, and Kurtosis.

These features when calculated for every one of the 19 features existing in the original dataset, give a total of 133 features. One tuple in the new dataset represents one file of the original dataset which consisted of 12,118 tuples. Each of these tuples thus provides a representative distribution of the data contained in the file it pertains to. The distribution can hence show the characteristics of the 12,118 tuples in a single tuple. This is essentially the data compression technique used to aid the previously substantial time required during the modeling process. The newly transformed dataset consists of 133 features and 310 tuples.

VGRF_left_s1Max	VGRF_left_s1Std	...	Total_force_leftAvg	Total_force_leftSkewness
480.92	140.246070	...	520.643543	-0.074995
495.55	134.295604	...	441.192562	-0.098336
451.88	133.190085	...	429.147671	-0.039759

Fig 6: Statistical Transformation of Dataset

Fig 6 is a small snippet of the data illustrating the new features of an original feature VGRF_left_s8. Finally, the

data is scaled between values (0,1) to improve overall consistency as can be seen in the Fig 7.

	0	1	2	...	130	131	132
0	0.0	0.352498	0.352497	...	0.128202	0.434056	0.347977
1	0.0	0.352498	0.352497	...	0.526895	0.278609	0.314254
2	0.0	0.352498	0.352497	...	0.216990	0.330556	0.058698
3	0.0	0.352498	0.352497	...	0.648126	0.382608	0.170110

Fig 7: Scaled version of the Transformed Dataset

VI. CLASSIFICATION TECHNIQUES

A total of 5 different classification algorithms have been used in an attempt to achieve accuracy as high as possible. K-Nearest Neighbors Classifier, Logistic Regression, SVM (Linear Kernel, RBF Kernel, Poly Kernel), Decision Trees, and Random Forest.

A. K-Nearest Neighbors

K-Nearest Neighbors, one of the simpler machine learning techniques, is a non-parametric method which uses similarity measure (distance function) to classify data based on the data it has already received. It accepts a parameter k which defines how many data points the classifier requires to make a prediction about a new data point. In other words, a case is classified by a majority vote of its neighbors, with the case being assigned to the class most common amongst its K nearest neighbors measured by a distance function.

The K-Nearest Neighbors model achieved the best accuracy of 93.08% at n_neighbors = 2 on the data. To achieve this accuracy, we iterated over different values of k (neighbors) and found out the mean cross-validated accuracy achievable using GridSearchCV in scikit-learn [3]. The precision, recall, and F1 scores calculated are 89.58%, 84.31%, and 86.86% respectively.

B. Decision Trees

Decision Trees, is a systematic approach which poses a series of carefully crafted questions or conditions to segregate between data based on their attributes. It accepts multiple parameters which can be manipulated to get higher accuracy.

Decision Tree Model was iterated over different values of max_depth which defines the height of the tree. When set to max_depth=100, it was able to achieve an accuracy of 87.8%. GridSearchCV [3] was able to give the mean cross-validated score for a range of values of the parameter. This also helps in deciphering the best value for that parameter, in this case, max_depth. The Precision, Recall and f1 score were 80.7%, 92.0%, and 85.98% respectively for this model.

C. Random Forests

Random Forests, in simpler terms, uses an ensemble of trees to make prediction.. It uses averaging to improve the predictive accuracy and control overfitting which is why it more generalised and tolerant than decision trees.

Random Forests model achieved an accuracy of 90.39%. When iterated over different values of n_estimators using GridSearCV [3], the optimal value of the paramter came out to be 40. n_estimators parameter governs how many trees should be taken into consideration at once to form an ensemble. 87.72%, 90.91%, 89.29% were the precision, recall and f1 scores respectively for decision trees.

D. Support Vector Machines

Tahrir et al. [7] showed that SVMs can diagnose Parkinson's from a combination of spatiotemporal, kinematic, and kinetic gait data. In that study, spatiotemporal data was collected using infrared sensors attached to the subjects' hips and legs, while kinetic data was collected using force sensors placed on the subjects' feet [9]. SVM model was implemented for 3 kernels: Linear, RBF, and Poly kernels. All these kernels input C parameter which governs how powerfully the training data fits on the model. There is another parameter gamma used in the case of poly and radial basis function (rbf) kernels. A good combination of these two parameters can bring about a model with high accuracy. Best combinations possible for these kernels were recorded using the GridSearchCV function in the scikit-learn [3]. Linear Kernel gave an accuracy of 89.95% for C=1. The Precision, Recall and f1 score were 81.13%, 87.86%, and 84.31% respectively for this kernel.

RBF kernel gave the best accuracy of 90.39% at C= 50 and gamma = 0.01. The Precision, Recall and f1 score were 85.45%, 90.38%, and 87.85% respectively for this kernel.

Poly kernel was able to yield an accuracy of 88.64% at C= 0.1, gamma = 1 and degree = 3. The Precision, Recall and f1 score were 87.93%, 94.44% and 91.07% respectively for poly kernel.

E. Logistic Regression

It is a binary classification algorithm used when the response variable is dichotomous (1 or 0). Inherently, it returns the set of probabilities of the target class.

An Accuracy of 90.06% has been achieved by using a Logistic Regression model. Using GridSearchCV over different values of C within a suitable range, we found that the mean cross-validated score for this model was 90.06% when C was set to 5. The precision, recall, and F1 scores calculated are 78.18%, 87.76%, and 82.69% respectively.

Since the scikit-learn's [3] train_test_split() function uses an attribute called random_state which will affect the accuracy value, the accuracies are bound to change a little each time this function is run before applying the models.

VII. RESULTS

Detection of Parkinson's disease using Gait analysis has been successfully performed using the Machine Learning Techniques mentioned previously. The accuracies obtained have been restated in Table 1. The parameters at which the following accuracies have been obtained are listed along with the algorithms and their accuracies. Another table which states the precision, recall and f1 scores of each of the models has been shown in Table 2.

It is to be noted that scikit-learn's train_test_split() function has been used for the purpose of splitting the transformed dataset into training and testing data. This function intelligently selects an equal number of data for training the model from each of the classes (the classes being 0 and 1 differentiating CO from PD patients). The function, however, employs a random number generator due to which accuracies vary a little every time it is run. Hence, a little variance in accuracies is expected when reproduction of the models is done.

Classifier	Accuracy	Parameters
Logistic Regression	90.06%	C = 5
SVM (Linear Kernel)	89.95%	C = 1
SVM (RBF Kernel)	90.39%	C = 50, gamma = 0.01
SVM (Poly Kernel)	88.64%	C = 0.1, gamma = 1, degree= 3
K-Nearest Neighbors	93.08%	K = 2
Decision Tree	87.78%	Max-Depth = 100
Random Forest	90.39%	n_estimators = 40

Table 1: Classifiers and their Accuracies with reproduction parameters

Classifier	Precision	Recall	F1 Score
Logistic Regression	78.18%	87.76%	0.8269
SVM (Linear Kernel)	81.13%	87.86%	0.8431
SVM (RBF Kernel)	85.45%	90.38%	0.8785
SVM (Poly Kernel)	87.93%	94.44%	0.9107
K-Nearest Neighbors	89.58%	84.31%	0.8686
Decision Tree	80.70%	92.00%	0.8598
Random Forest	87.72%	90.91%	0.8929

Table 2: Precision, Recall and F1 scores

VIII. SIGNIFICANCE OF RESULTS

Statistical Compression Techniques have modeled better classification of the data as compared to the traditional Principal Component Analysis methodology. This is also because PCA is not applicable as a sound technique for dimensionality reduction when the data is not in the form of a Gaussian Distribution. This is one of the key reasons PCA failed to provide better results as compared to the Statistical Techniques.

Finally, the detection of Parkinson's disease is a highly divisive issue and we try to incorporate parameter values which provide us with a higher value of the recall. This is because, in highly sensitive environments like disease detection, the accuracy measure may not always be the best parameter to judge our model on. The false positives do not cause as much harm as do the false negatives.

IX. CONCLUSION

In this project, we presented a way to classify PD patients from Control Objects using Gait Analysis Data. Through Univariate Analysis of data, we transformed it by applying Statistical techniques which represented its overall distribution, thus the data was significantly reduced. Prediction models were then applied to the newly transformed data which were successfully able to classify between Control Objects and PD patients. These models can be used to test for PD patients by analyzing their VGRF data of stride patterns as done in gait analysis.

In addition to that, we prove that often times, for data with tons of tuples and comparatively fewer features, statistical analysis techniques can be much more helping rather than reliance on Principal Component Analysis (PCA). Since PCA can only reduce the dimensionality of the data in a feature space sense, the number of tuples remain the same, which leads to a high computational power requirement in case the tuples are in millions. Statistical data analysis techniques supersede PCA here because incorporating them we could find a trade-off between the number of tuples and the feature space. An extensive data compression has been carried out using these techniques, henceforth reducing the number of tuples by 99.99065489% (since 3317244 tuples were reduced into 310 tuples) while the feature space was increased by 600% (since 19 tuples were increased to 133 features).

REFERENCES

[1] Understanding Parkinson's Disease, Parkinson's Disease Foundation, 2014

[2] Mazzoni P, Shabbott B, Cortés JC. Motor control abnormalities in Parkinson's disease. *Cold Spring Harb Perspect Med* 2012.

[3] Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.

[4] Yogev G, Giladi N, Peretz C, Springer S, Simon ES, Hausdorff JM. Dual tasking, gait rhythmicity, and Parkinson's disease: Which aspects of gait are attention demanding? *Eur J Neuroscience* 2005; 22:1248-1256.

[5] Frenkel-Toledo S, Giladi N, Peretz C, Herman T, Gruendlinger L, Hausdorff JM. 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* 2005: 2:23.

[6] Matplotlib: A 2D Graphics Environment, J.D. Hunter *et al.,* Computing in Science & Engineering, IEEE Computer SOC.

[7] N. Tahir, H. Manap. Parkinson Disease Gait Classification based on Machine Learning Approach, Journal of Applied Sciences 2012, 12(2), 180-185.

[8] Salarian A, Russmann H, Vingerhoets FJ, Dehollain C, Blanc Y, Burkhard PR, Aminian K: Gait assessment in Parkinson's disease: toward an ambulatory system for long-term monitoring. IEEE Trans Biomed Eng 2004, 51: 156-159.

[9] Chang, D., Alban-Hidalgo, M., Hsu, K. (2014). Diagnosing Parkinson's disease from gait. *Stanford*.